**UNIVERSIDADE DE SANTIAGO DE COMPOSTELA**

**Departamento de Xenética**

# Evolución de los Retrotransposones con LTRs del Grupo *Ty3/gypsy* en los Genomas de *Anopheles gambiae*, *Aedes aegypti* y *Drosophila melanogaster*.

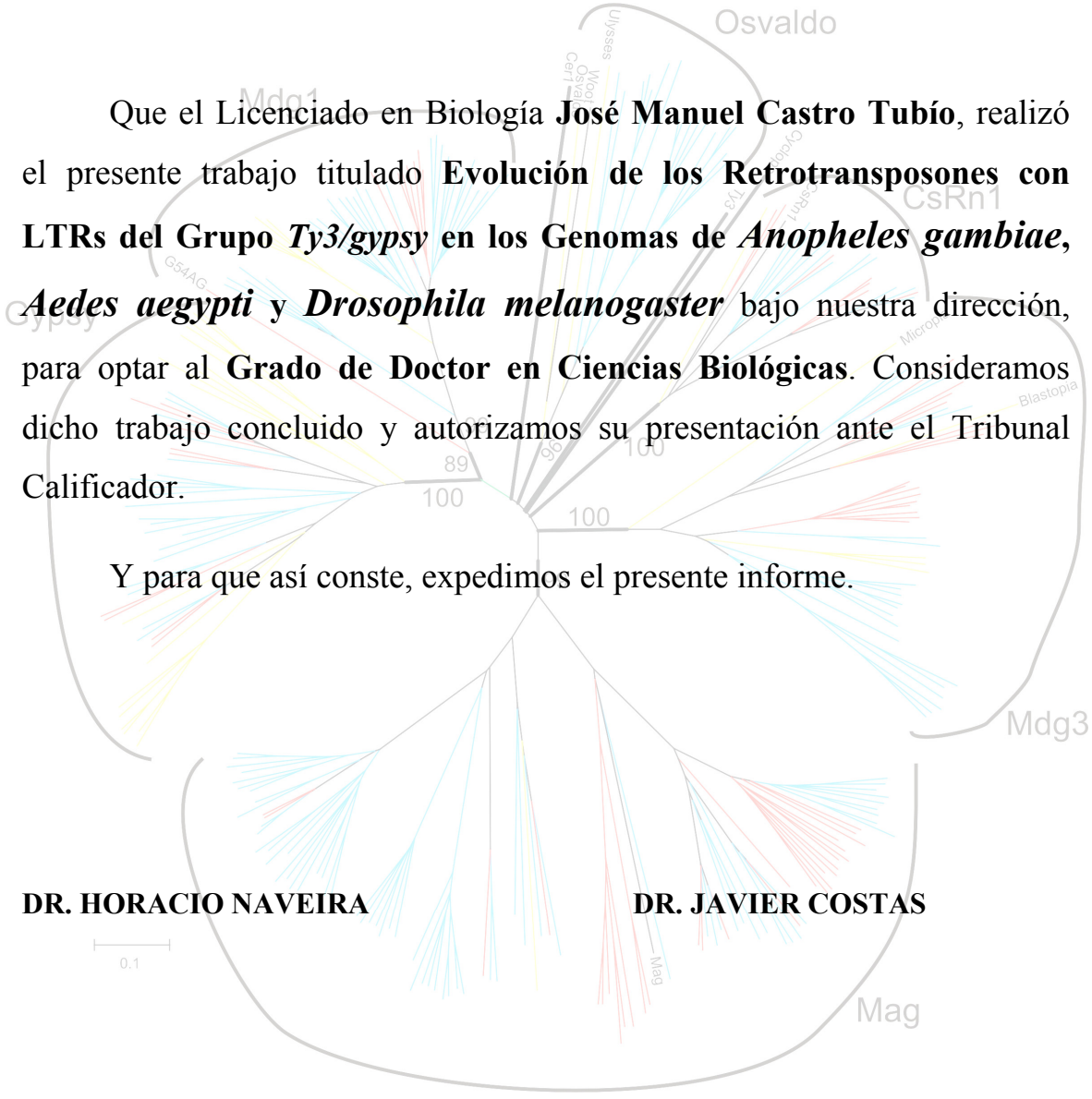
**TESIS DOCTORAL**

**JOSÉ MANUEL CASTRO TUBÍO**

Mayo, 2009

LOS DOCTORES **EMILIO VALADÉ DEL RÍO**, **HORACIO NAVEIRA FACHAL** Y **JAVIER COSTAS COSTAS**,

**INFORMAMOS**




Que el Licenciado en Biología **José Manuel Castro Tubío**, realizó el presente trabajo titulado **Evolución de los Retrotransposones con LTRs del Grupo *Ty3/gypsy* en los Genomas de *Anopheles gambiae*, *Aedes aegypti* y *Drosophila melanogaster*** bajo nuestra dirección, para optar al **Grado de Doctor en Ciencias Biológicas**. Consideramos dicho trabajo concluido y autorizamos su presentación ante el Tribunal Calificador.

Y para que así conste, expedimos el presente informe.

**DR. HORACIO NAVEIRA** **DR. JAVIER COSTAS**

**DR. EMILIO VALADÉ** **El doctorando**

Santiago de Compostela, 20 de Mayo de 2009.

# I. Introducción

## I.1. *Anopheles gambiae*.

*Anopheles gambiae* es el principal vector del protozoo *Plasmodium falciparum* en África y es uno de los vectores de la Malaria más eficientes en el mundo. La Malaria, considerada la enfermedad parasítica más importante en el mundo, es responsable de 500 millones de enfermos y de hasta 2.7 millones de muertes cada año, más del 90% de los cuales ocurren en el África subsahariana (Breman, Egan & Keusch, 2001).

Por todos nosotros es conocido, en mayor o menor grado, el interés sanitario que supondría disponer de un mayor conocimiento sobre cualquier aspecto relacionado con la biología de este insecto y, teniendo en cuenta que nos encontramos en la ya denominada Era de la Genómica, era de esperar que el genoma de *Anopheles gambiae* fuera uno de los primeros genomas eucariotas en ser secuenciado completamente (Holt et al., 2002).

### I.1.1. Aspectos generales acerca del mosquito *Anopheles gambiae*.

El género *Anopheles* es con mucho el más amplio de los tres géneros que comprende la subfamilia *Anophelinae*, el linaje más basal del grupo de los Mosquitos (Reino *Animalia*, Filo *Arthropoda*, Clase *Insecta*, Orden *Diptera:* Familia *Culicidae*). El género está representado por cerca de 500 especies que pueden encontrarse en todos los continentes, excepto la Antártida. Entre los insectos, los anofelinos son estudiados debido a su importancia médica, ya que actúan como vectores de *Plasmodium* (protozoo causante de la Malaria), microfilarias (nematodos causantes de la Filariasis) y virus del tipo de los *Arbovirus*.

El vector más importante del parásito *Plasmodium*, causante de la Malaria, en el África subsahariana es el mosquito *Anopheles gambiae sensu stricto* (*Anopheles gambiae s. s.*). Este mosquito es clasificado dentro de un grupo de especies estrechamente emparentadas conocido como "Complejo de *Anopheles gambiae*". Las especies de este Complejo se caracterizan por ser casi indistinguibles morfológicamente.

*Anopheles gambiae s. s.* muestra una extrema heterogeneidad genética. Por una parte los polimorfismos de inversión cromosómica llevaron a la identificación de hasta

cinco "formas cromosómicas" y, por otra, el descubrimiento de diferencias a nivel del ADN ribosomal (rDNA) ha permitido concluir que dentro de la especie existen dos taxones en pleno proceso de especiación y, por tanto, con restricciones al flujo génico entre ambos (Krzywinski & Besansky, 2003).

### I.1.1.1. Sistemática superior de la subfamilia *Anophelinae*.

Tradicionalmente, la subfamilia *Anophelinae* se subdivide en tres géneros: *Anopheles*, *Bironella* y *Chagasia*. Dado el elevado número de especies que lo conforman, el género *Anopheles* fue a su vez subdividido en los subgéneros *Anopheles*, *Cellia*, *Kerteszia*, *Lophopodomyia*, *Nyssorhynchus* y *Stehomyia* (Krzywinski & Besansky, 2003). Según Krzywinski & Besansky (2003), de todos los estudios relativos a las relaciones filogenéticas dentro de la subfamilia *Anophelinae*, se pueden tomar con seguridad las siguientes conclusiones (ver figura 1): 1) todos los estudios sugieren que *Chagasia* fue el primer linaje en radiar del árbol de los anofelinos; 2) el gen *white* predice que la siguiente divergencia corresponde a la de *Brionella* y el género *Anopheles*; 3) dentro del género de *Anopheles* la rama más basal corresponde a *Stethomyia* y, aunque la filogenia es poco clara, en los otros grupos las evidencias sugieren la radiación de dos clados hermanos: por un lado *Cellia* con el subgénero *Anopheles* y por el otro *Lophopodomyia* con *Kerteszia* y *Nyssorhynchus*.

La filogenia dentro del subgénero *Anopheles* abarca varios grupos de especies y "Complejos". Al menos la mitad de los vectores importantes de la Malaria se constituyen en complejos de especies muy estrechamente relacionadas cuyos miembros son isomórficos o, al menos, muy similares (Collins & Paskewitz, 1996). Esta relación genética tan próxima complica mucho la resolución de relaciones filogenéticas por dos razones: la primera es que las especies comparten polimorfismo ancestral; la segunda es que las barreras reproductivas postapareamiento son incompletas, de manera que la esterilidad híbrida de la F1 suele afectar a los machos pero no a las hembras, permitiendo a estas últimas actuar como "puertas" a la introgresión y, por lo tanto, al flujo génico entre especies.

El vector afrotropical *Anopheles gambiae sensu stricto* es una de las, al menos, siete especies estrechamente emparentadas y morfológicamente indistinguibles que

conforman el denominado "Complejo de *Anopheles gambiae*" (Davidson, 1964). Actualmente, las especies reconocidas como parte del Complejo son *Anopheles melas*, *Anopheles merus*, *Anopheles bwanbae*, *Anopheles quadriannulatus* A, *Anopheles quadriannulatus* B, *Anopheles arabiensis* y, por supuesto *Anopheles gambiae*. Además, los descubrimientos recientes acerca de la estructuración de las poblaciones de esta última especie prometen su separación en otras dos, que de momento se conocen como formas moleculares *M* y *S* de *Anopheles gambiae sensu stricto*. También se han detectado procesos de especiación incipiente en *Anopheles arabiensis* y *Anopheles melas*.

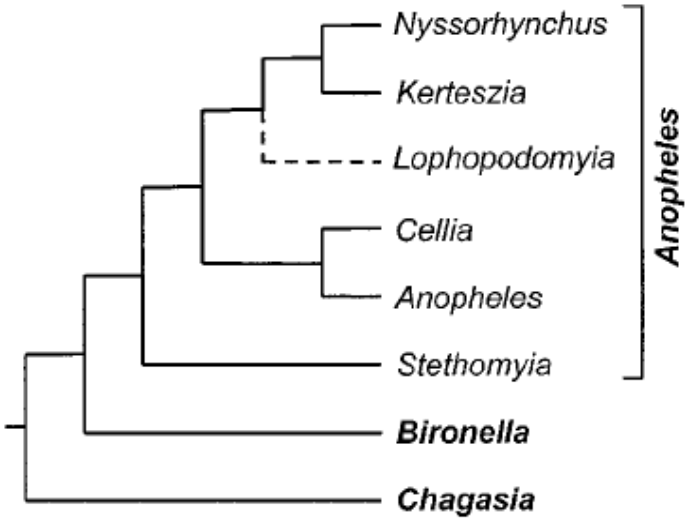



**Figura 1.** Relaciones filogenéticas dentro de la subfamilia *Anophelinae*. Tomada de Krzywinski & Besansky (2003).

### I.1.1.2. Biogeografía de los anofelinos.

Dentro del género *Anopheles*, los subgéneros *Kerteszia*, *Lophopodomyia*, *Nyssorhynchus* y *Stethomyia* ocupan América del Sur, *Cellia* se encuentra en África y el subgénero *Anopheles* es cosmopolita. La posición filogenética basal del género neotropical *Chagasia* (figura 1) y la distribución, también neotropical, de cuatro de los seis subgéneros del género *Anopheles*, lleva a la conclusión de que la subfamilia *Anophelinae* se habría originado en el Nuevo Mundo (Krzywinski, Wilkerson & Besansky, 2001).

Se ha propuesto que las primeras divergencias dentro de la subfamilia *Anophelinae* tuvieron lugar mucho antes de la disgregación de Gondwana (Krzywinski, Wilkerson & Besansky, 2001), y que las primeras radiaciones dentro del subgénero *Anopheles* debieron haber tenido lugar antes de que se perdiera la conexión entre África y Sudamérica (hace unos 95 millones de años). Los puentes de tierra que mantenían unidos a África y Europa (creados en el Paleoceno) y la conexión desde Europa hasta Norteamérica (existente hasta el final del Eoceno) permitieron una mayor dispersión del subgénero *Anopheles* dentro de Laurasia, y algunos de sus linajes debieron entran más tarde en Sudamérica desde el Norte.

Existen unos interesantes paralelismos entre la hipótesis arriba comentada y la filogenia propuesta para los mamíferos placentarios, que seguramente también se hayan originado en Gondwana y, como en los anofelinos, sus linajes basales experimentaron una rápida diversificación, probablemente coincidiendo con la separación de África y Sudamérica. Bajo este escenario, resulta fácil concebir que la radiación de los anofelinos fuera consecuencia de la radiación de sus hospedadores.

### I.1.1.3. Distribución geográfica de las especies del Complejo de *Anopheles gambiae*.

Dentro del Complejo de *Anopheles gambiae*, las especies *A. gambiae* y *A. arabiensis* son los dos miembros más antropofílicos[a] y que abarcan un mayor rango de distribución, por lo que tienen una importancia médica principal. Puede afirmarse, de una manera general, que la especie *gambiae* predomina en zonas forestales y de sabana húmeda, mientras que *arabiensis* predomina en las sabanas áridas y en zonas esteparias. Es importante señalar que *A. arabiensis*, a diferencia de *A. gambiae*, se encuentra ampliamente distribuida en el "cuerno" de África (territorio ocupado por Somalia y la parte más oriental de Etiopía) y es la única especie del Complejo presente en la península arábiga, ocupando una pequeña fracción en el Sudoeste peninsular. En aquellas áreas en las que *gambiae* y *arabiensis* presentan una distribución simpátrida y en donde la reproducción puede ocurrir a lo largo de todo el año, se ha observado que existen diferencias en cuanto a las estaciones preferentes para la reproducción por parte de ambas

[a] En parasitología, se entiende por *Antropofília* a la apetencia selectiva de ciertos artrópodos por la sangre humana.

especies, observándose un incremento en la frecuencia relativa de *arabiensis* en la estación seca. Otros datos han mostrado que *arabiensis* es más zoofílica y exofílica[b] que *gambiae*.

Las especies *Anopheles quadriannulatus* A y B son marcadamente zoofílicas, por lo que tienen una menor importancia médica. Ambas especies presentan una distribución simpátrida con *A. arabiensis* y, en menor grado, con *A. gambiae*. Parecen tener una distribución restringida a África Sudoriental, Etiopía y Zanzíbar. En Zanzíbar y el Sur de África *A. quadriannulatus* parece ser completamente exofílica, mientras que tiende hacia la endofília en Etiopía. Estas dos especies, de distribución alopátrida, representan reliquias de una especie *A. quadriannulatus* ancestral, y llegarían a diferenciarse genéticamente tras el aislamiento geográfico de ésta.

Por último, *Anopheles bwanbae*, *Anopheles melas* y *Anopheles merus* son las especies del Complejo con menor distribución. *A. bwanbae* presenta una distribución geográfica restringida al Noreste de Uganda, *A. melas* se distribuye a lo largo de la costa Oeste africana y *A. merus* a lo largo de la costa Este, adentrándose en Sudáfrica. Además, esta última especie ha sido recientemente identificada en Madagascar (Leong et al., 2003), isla donde ya era conocida la presencia de otras dos especies del Complejo: *A. gambiae* y *A. arabiensis*.

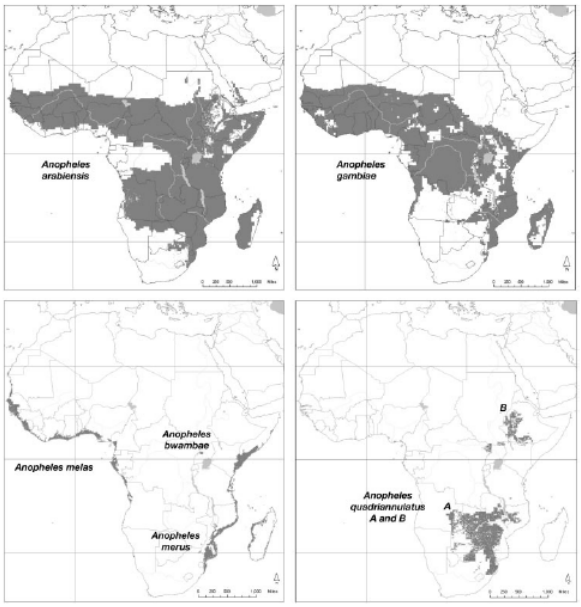



**Figura 2.** Distribución geográfica de *Anopheles gambiae* y las otras seis especies estrechamente emparentadas del Complejo. Tomada de Ayala & Coluzzi (2005).

[b] Entiéndase por *Endofília* y por *Exofília* al hábito de los mosquitos por reposar dentro o fuera de aquellas construcciones humanas donde se realizan las actividades cotidianas.

### I.1.1.4. Análisis del complemento politénico en las especies del Complejo de *A. gambiae*.

Los reordenamientos cromosómicos pueden desempeñar un importe papel en la evolución de las especies (Coghlan et al., 2005). En el caso de los anofelinos, la sintenia[c] se encuentra altamente conservada, pero el orden de los genes ha sido extensivamente cambiado, principalmente a través de inversiones cromosómicas paracéntricas (Cornel & Collins, 2000; Sharakhov et al., 2002). Debido a la supresión recombinacional entre reordenamientos alternativos y a la estabilización de determinadas combinaciones con valor adaptativo, las inversiones paracéntricas también han desempeñado un papel importante en la especiación de las especies del Complejo de *Anopheles gambiae* (Coluzzi et al., 2002; Ayala & Coluzzi, 2005).

El estudio del complemento politénico dentro del Complejo de *A. gambiae* ha resultado fundamental para la identificación de las diferentes especies morfológicamente indistinguibles, o casi morfológicamente indistinguibles, que lo conforman. Al igual que en casi todos los mosquitos, las especies de este Complejo presentan un cariotipo mitótico de dos pares de autosomas y un par de cromosomas sexuales. El complemento politénico consiste en cinco brazos cromosómicos (*X*, *2R*, *2L*, *3R*, *3L*), siendo el patrón de bandas claro y, además, su correspondencia entre las distintas especies del Complejo resulta fácilmente identificable, con la experiencia adecuada, exceptuando las comparaciones del cromosoma *X* entre algunas especies.

En las especies del Complejo las inversiones cromosómicas paracéntricas son abundantes, representando éstas los cambios esenciales que diferencian el patrón de bandas entre distintas especies. Estos reordenamientos fueron clasificados en relación a la secuencia politénica de la especie *Anopheles quadriannulatus*. En las diferentes especies del Complejo se han identificado diez inversiones fijadas y hasta más de 120 inversiones polimórficas (Coluzzi et al., 1979; Coluzzi et al., 2002). La mayoría de las inversiones paracéntricas se localizan a lo largo del cromosoma *2R*. El análisis de estas inversiones ha venido siendo utilizado en la identificación de las especies del Complejo, así como de individuos. Además, el estudio de estas inversiones también ha sido empleado en la interpretación de la historia evolutiva de las especies del Complejo.

[c] El término *Sintenia* se refiere a la presencia de genes homólogos en un mismo cromosoma, al comparar entre si diferentes especies con un antepasado común.

*A. gambiae* y *A. arabiensis* han desarrollado un patrón muy complejo de polimorfismos de inversión cromosómica, representando las especies del Complejo con mayor número de estos polimorfismos, seguidas por *A. melas*. Sin embargo, se conocen muy pocos polimorfismos en *A. bwambae* y *A. quadriannulatus* especie *A*, mientras que no se conocen polimorfismos de inversión en *A. quadriannulatus* especie *B* y *A. merus*.

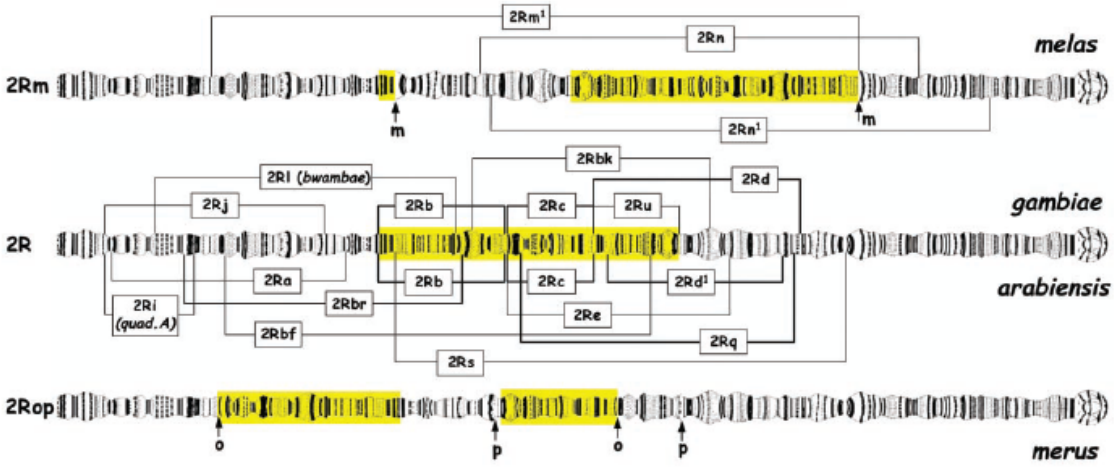



**Figura 4.** Principales inversiones paracéntricas del cromosoma *2R* en las especies del Complejo de *Anopheles gambiae*. Los reordenamientos fijados caracterizan al cromosoma *2R* de *A. melas* (*2Rm*, donde ↑m indica los límites de la inversión) y de *A. merus* (*2Rop*, donde ↑o y ↑p marcan los límites de las inversiones solapadas *o* y *p*). Los corchetes marcan los límites de las inversiones polimórficas en *A. gambiae*, *A. arabiensis*, *A. melas* y *A. merus*. Tomada de Coluzzi et al. (2002).

Asumiendo una ocurrencia aleatoria de los puntos de ruptura de estas inversiones, el número esperado de inversiones en cada cromosoma politénico dependería de las longitudes de los cromosomas. Sin embargo, en el cromosoma *X* se localizan la mitad de las 10 inversiones fijadas del complejo, a pesar de que este cromosoma solamente representa el 11% del total del complemento politénico, mientras que el cromosoma *3* solamente contiene una inversión fijada, a pesar de que representa un 37% del complemento. El número esperado de inversiones fijadas se encuentra solamente en el cromosoma *2*. Sin embargo, el 58% de las inversiones polimórficas (18/31) se localizan en el brazo *R* del cromosoma *2*, a pesar de que este brazo solamente representa menos del 30% del complemento politénico total. Según Coluzzi et al (1979, 2002), esta distribución no aleatoria de las inversiones sugiere que la localización de estos reordenamientos es

producto de la acción de la selección natural, de manera que, tal y como se explica más adelante en el texto, las inversiones cromosómicas constituirían un mecanismo para la diferenciación ecotípica[d]. Además, esta hipótesis se ve reforzada en cuanto a que los puntos de ruptura de las inversiones no se distribuyen de forma uniforme a lo largo del cromosoma *2R*. De hecho, se sabe que uno de los puntos de ruptura en el cromosoma *2R* es el mismo en tres inversiones (*c*, *d* y *u*), o al menos coinciden citológicamente. Un caso similar, muy estudiado, es el del grupo *repleta* de *Drosophila*, en el que una parte desproporcionada de las inversiones se localiza en el cromosoma *2*.

[d] Un *Ecotipo* es una subpoblación genéticamente diferenciada que está restringida a un hábitat específico, un ambiente particular o un ecosistema definido.

### I.1.2. La especiación en los mosquitos del Complejo de *Anopheles gambiae*.

#### I.1.2.1. Las relaciones filogenéticas entre las especies del Complejo.

La verdadera topología de las relaciones filogenéticas dentro del Complejo de *Anopheles gambiae* resulta ser muy complicada de discernir, ya que el Complejo de *A. gambiae* comprende un grupo de especies estrechamente relacionadas, que han divergido recientemente y en un período tiempo muy corto. De hecho, la distancia genética de Nei entre las especies que conforman el Complejo fue estimada en un rango de tan solo 0.1-0.25 (Coluzzi, Petrarca & DiDeco, 1985). Como comparación, sirvan los valores de este índice obtenidos para las especies sinmórficas de los Complejos *buzzatii* y *willistoni* de *Drosophila*, que resultaron 0.59 y 0.58, respectivamente (Fontdevila & Moya, 2003). Además, experimentos sobre el cruzamiento interespecífico en laboratorio han determinado que el aislamiento reproductivo es incompleto, habiéndose encontrando híbridos en la naturaleza tan solo ocasionalmente (Besansky et al, 1994).

Tradicionalmente, se ha venido aceptando la filogenia que sitúa a *A. gambiae* y *A. arabiensis* en extremos opuestos del árbol filogenético. Esta relación (figura 5) se ha extraído mediante un análisis parsimonioso de las inversiones paracéntricas identificadas en los cromosomas politénicos de las especies del Complejo (Coluzzi et al., 1979). Según Coluzzi y sus colaboradores, en el Complejo de *A. gambiae* hay dos inversiones sinapomórficas[e]: la inversión *ag* del cromosoma *X* (denominada inversión $X^{ag}$) y, por otra parte, la inversión *a* del cromosoma *3L* ($3L^{a}$). La primera de estas inversiones indicaría que *A. gambiae* y *A. merus* son taxones hermanos y, por otra parte, la inversión $3L^{a}$ ligaría de la misma forma a *A. melas* y *A. bwambae*. Estas relaciones filogenéticas fueron propuestas en base a que ambas inversiones ($X^{ag}$ y $3L^{a}$) serían monofiléticas. Los autores eran conscientes de que, asumiendo el origen monofilético de ambas inversiones, deberían explicar la distribución de las inversiones $2R^{b}$, $2R^{bc}$ y $2L^{a}$, y es que los polimorfismos $2R^{b}$ y $2R^{bc}$ son compartidos por *A. gambiae* y *A. arabiensis* y, por otra parte, la inversión $2L^{a}$ es polimórfica en *A. gambiae* y está fijada en *A. arabiensis* y *A. merus*. Además, el análisis mediante microscopía óptica de los cromosomas politénicos y el hecho de que aparean perfectamente en híbridos, indicaba que estas inversiones eran idénticas en *A. gambiae* y *A. arabiensis*. Todo esto llevó a Coluzzi et al. (1979) a proponer introgresión vía hembras

[e] *Sinapomorfismo* hace referencia a un carácter homólogo apomórfico (es decir, una novedad evolutiva) compartido por dos o más taxones.

híbridas fértiles. Además, consideraron también, como otra posible explicación, que se haya mantenido un polimorfismo ancestral a lo largo de los diferentes procesos de especiación, aunque esta posibilidad sería la menos probable debido a que frecuentemente los procesos de especiación están asociados a cuellos de botella y, además, ya se había constatado que existían en la naturaleza híbridos entre *A. gambiae* y *A. arabiensis*, lo que sin duda apoyaría la propuesta de la introgresión más que esta última opción.

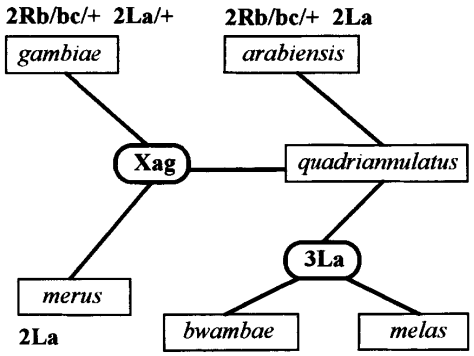



**Figura 5.** Esquema que muestra la proximidad filogenética entre las especies del Complejo de *A. gambiae*, y que está basada en las inversiones compartidas entre las mismas. Este esquema, recogido en della Torre et al. (1997), viene a ser un resumen del que originariamente había sido elaborado por Coluzzi et al. (1979).

Las relaciones propuestas por Coluzzi violaban, sin embargo, los datos morfológicos, etológicos, ecológicos, y los datos que existían acerca de la hibridación entre las especies del Complejo. Para testar la veracidad de las predicciones de Coluzzi, sobre la filogenia basada en las inversiones cromosómicas, Besansky et al. (1994) obtuvieron las secuencias procedentes de dos cromosomas diferentes (secuencias del ADN ribosomal y del gen de la esterasa) y del ADN mitocondrial en 5 taxones del Complejo de *A. gambiae*. Tras obtener las filogenias correspondientes (por máxima parsimonia, *neighborg joining* y *maximum likelihood*) se veía reforzada la relación de *A. gambiae* y *A. arabiensis* como taxones hermanos (figura 6). Los autores concluían que la filogenia de Coluzzi, basada en las inversiones cromosómicas, no era la acertada para las relaciones entre las especies del Complejo, sugiriendo que las relaciones propuestas por ésta última estarían influenciados por introgresiones dentro del Complejo y a que el origen de dichas inversiones sería polifilético o parafilético. Sin embargo, las discrepancias no habían hecho más que empezar, y a lo largo de la siguiente década se han ido sucediendo diversas investigaciones con la intención de revelar la verdadera filogenia dentro del Complejo.

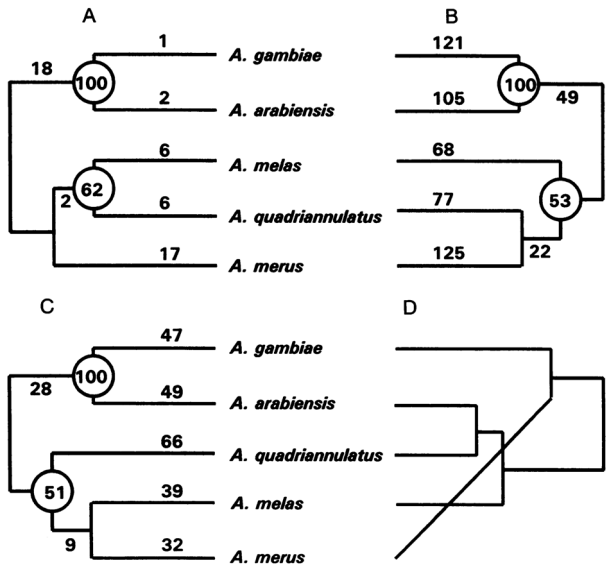



**Figura 6.** Árboles de Máxima Parsimonia obtenidos por Besansky et al. (1994), basados en el alineamiento de secuencias del DNA mitocondrial (A), DNA ribosómico (B), combinación del DNA mitocondrial y del DNA ribosómico (C) y basado en las inversiones cromosómicas asumiendo un origen monofilético de las inversiones (D). Los números dentro de los círculos son valores *bootstrap* basados en 100 réplicas.

No fue solamente el trabajo de Besansky el que había dado con datos que cuestionaban la veracidad de la filogenia de Coluzzi, sino que también otro trabajo publicado dos años después (Caccone, Garcia & Powell, 1996) ponía en entredicho ciertas relaciones marcadas por las inversiones, concluyendo que *A. arabiensis* y *A. gambiae* serían los verdaderos taxones hermanos, excluyendo a *A. merus* de su relación próxima a *A. gambiae*. Caccone et al. (1996) secuenciaron la región control rica en *AT* del ADN mitocondrial en seis especies del Complejo de *A. gambiae*, encontrando conformidad con las relaciones filogenéticas sugeridas por las inversiones cromosómicas, excepto en un punto: se reforzaba la idea de *A. gambiae* y *A. arabiensis* como taxones hermanos, frente a *A. gambiae* y *A. merus*. Los autores comentaban que la posición filogenética anómala de *A. arabiensis* podía ser debida a introgresión con *A. gambiae*. Además, concluían la existencia de flujo génico, a nivel de ADN mitocondrial, entre *A. gambiae* y *A. arabiensis*.

Además de la introgresión, otra explicación sencilla para revelar la verdadera topología entre las relaciones *A. gambiae*/*A. arabiensis* frente a *A. gambiae*/*A. merus*, podría ser que la inversión $X^{ag}$ no sea monofilética, de forma que su presencia en *A.*

*gambiae* y en *A. merus* podría deberse a que la inversión surgió más de una vez en procesos totalmente independientes, o bien a que lo que se denomina $X^{ag}$ en las dos especies no es realmente la misma inversión, algo que no se puede saber con el análisis rutinario de los cromosomas politénicos a través de un microscopio de luz, sino que se haría necesario el estudio a nivel molecular de los extremos de dichas inversiones para compararlos en las diferentes especies implicadas. Así, García et al. (1996), tras estudiar el gen de la guanilato ciclasa localizado en el interior de la inversión $X^{ag}$, propusieron que dicha inversión tenía un origen monofilético para *A. gambiae* y *A. merus* (figura 7). El razonamiento que dio origen a este estudio fue el siguiente: dado que las inversiones protegen de la recombinación a los genes que contienen, si la inversión es monofilética las secuencias de ADN contenidas en la inversión deberían reflejar la verdadera historia evolutiva de la inversión, porque un gen encerrado en el interior de una inversión se mantendrá ligado completamente a esa inversión, salvo por doble entrecruzamiento: hay evidencias de que el tamaño de las inversiones que se observan en la naturaleza resulta de un compromiso entre la ventaja de combinar loci alejados en un bloque coadaptado y la desventaja de que a mayor longitud del segmento invertido, menor restricción de la recombinación (Caceres et al., 1999). La idea de García y sus colaboradores venía abalada por trabajos previos que estudiaron a *Drosophila* (Aquadro et al., 1991; Popadic & Anderson, 1994; Rozas & Aguade, 1994), en donde la monofilia de determinadas inversiones había sido confirmada por el hecho de que secuencias de ADN contenidas en las inversiones producían árboles filogenéticos idénticos a los deducidos por las inversiones y, además, todos los alelos contenidos en la misma inversión eran más similares entre ellos que a cualquier otro alelo.

De acuerdo con este último trabajo, todo parecía indicar que la introgresión mediante hibridación entre especies podría ser la explicación para la distribución de las inversiones $2R^b$, $2R^{bc}$ y $2L^a$ en el Complejo, posibilidad que, tal y como se ha mencionado, ya había sido propuesta en el trabajo original de Coluzzi (1979). Con la intención de testar esta posibilidad, della Torre et al. (1997) llevaron a cabo cruzamientos experimentales en laboratorio entre las especies *A. gambiae* y *A. arabiensis*, observando que cuando se introgresaba el cromosoma *X* al cabo de dos generaciones no quedaba ningún híbrido con el cromosoma *X* introgresado. Esto no ocurría siempre que las introgresiones afectaban al cromosoma *2*, de manera que algunos de los cromosomas *2* introgresados perduraban a lo

largo del estudio de cruzamientos. Estos resultados mostraban que existe una elevada acción de la selección en contra de las introgresiones del cromosoma *X*, mientras que los híbridos con introgresiones del cromosoma *2* pueden persistir. Según los autores, recordando algunas de las predicciones de Coluzzi et al. (1979 y 1985): a través de la introgresión, *A. gambiae* adquiriría las inversiones $2R^b$ y $2L^a$ de *A. arabiensis*, lo que habilitaría a *gambiae* a expandirse desde los bosques húmedos (su hábitat original) hacia zonas más secas. De esta manera se veía reforzada la idea inicial propuesta en la filogenia de Coluzzi et al. (1979).

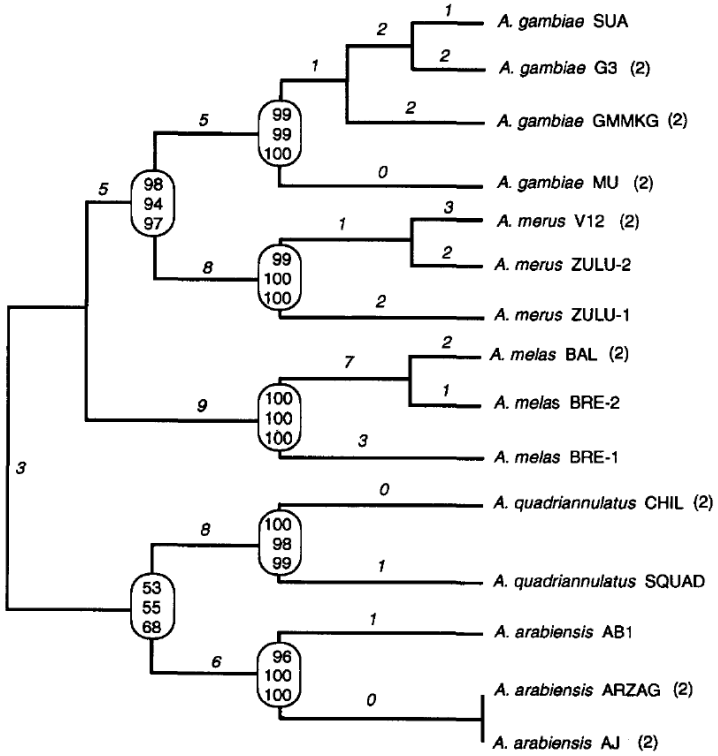



**Figura 7.** Árbol consenso *bootstrap*, obtenido por García et al. (1996), basado en el alineamiento de secuencias de ADN internas a la inversión $X^{ag}$ correspondientes al gen de la guanilato ciclasa. El árbol no tiene raíz. Los números en los nodos son los valores bootstrap (100 réplicas) para máxima parsimonia (arriba), *maximum likelihood* (en medio) y *neighbor-joining* (abajo). Figura tomada de García et al. (1996).

Si la filogenia proporcionada por Coluzzi et al. (1979) representa la verdadera historia evolutiva de las especies que conforman el Complejo de *A. gambiae*, es realmente complicado dar explicación a la distribución de la inversión $2L^a$. El trabajo de della Torre et al. (1997) ayuda a dar explicación a parte de la distribución de esta inversión en el Complejo. La inversión $2L^a$ está fijada en *A. merus* y en *A. arabiensis*, sin embargo, es polimórfica en *A. gambiae* (en donde se puede encontrar, además, el reordenamiento $2L^+$) y en el resto de las especies del Complejo es el reordenamiento standard $2L^+$ el que se

encuentra fijado. A partir de este punto, averiguar si la inversión $2L^a$ tiene realmente un origen único (monofilia) es fundamental para confirmar la verdadera historia evolutiva del Complejo. Si la inversión $2L^a$ fuera monofilética, entonces se pueden proponer dos hipótesis: 1) que existiera un polimorfismo ancestral $2L^a/2L^+$, de manera que con el tiempo estos reordenamientos se fijaran en unas especies u otras al tiempo que permanecería polimórfico en *A. gambiae*; 2) que la inversión $2L^a$ surgiera en una de las especies del Complejo y se transmitiera vía introgresión al resto de especies del complejo. En el caso de que la inversión $2L^a$ no fuera monofilética, entonces podría haber surgido en diferentes especies independientemente, o incluso podría ser que las diferentes inversiones $2L^a$ no fueran exactamente las mismas a nivel molecular, aunque fueran consideradas las mismas a través de la resolución de la microscopía óptica.

Para averiguar el verdadero origen de la inversión $2L^a$, Caccone, Min & Powell (1998) secuenciaron tres fragmentos de ADN interiores a la inversión $2L^a$, y uno exterior cercano a uno de los puntos de rotura, en individuos pertenecientes a diversas cepas de cinco de las especies del Complejo. La figura 8 muestra el resultado obtenido. Los autores concluyen que la explicación más parsimoniosa sobre la historia evolutiva de la inversión $2L^a$ es que el linaje común a las especies *A. merus* y *A. gambiae* tendría el reordenamiento $2L^+$ y que tras la divergencia de estas dos especies se generaría una inversión $2L^{a'}$ en *A. merus*, diferente a la inversión $2L^a$ que existe en *A. arabiensis*, pero que bajo el microscopio electrónico no puede diferenciarse de la inversión $2L^a$ de *A. arabiensis*. Esta inversión $2L^{a'}$ se fijaría en la especie *A. merus*. La introgresión de material genético desde *A. arabiensis* hacia *A. gambiae*, explicaría la adquisición de la inversión $2L^a$ original por parte de *A. gambiae* y, por tanto, el polimorfismo $2L^a$ /$2L^+$ hallado actualmente en esta especie. La introgresión de $2L^a$ desde *A. arabiensis* a *A. gambiae*, ya había sido propuesta años atrás por Coluzzi et al (1985) basándose en el patrón ecológico y biogeográfico de las especies. Es más, tal y como ya se ha mencionado, della Torre et al (1997) habían observado en laboratorio que las frecuencias de los híbridos *gambiae* $2L^+$/*arabiensis* $2L^a$ persistían en todas las generaciones durante la duración del experimento.

Tras la secuenciación del genoma de *Anopheles gambiae*, el grupo de Sharakhov (Sharakhov et al., 2006) llevó a cabo un estudio más detallado a nivel molecular, en donde se obtuvieron las secuencias de los puntos de rotura de la inversión $2L^a$ en tres especies del Complejo (*A. gambiae*, *A. arabiensis* y *A. merus*). Los datos obtenidos por estos

investigadores indican que el reordenamiento $2L^a$ es exactamente idéntico en las tres especies, rompiendo absolutamente las conclusiones obtenidas previamente por Caccone et al. (1998). Según Sharakhov y sus colaboradores, tal vez la conversión génica o la recombinación entre reordenamientos alternativos a alguna distancia de los puntos de rotura sea la causa de los resultados obtenidos por el grupo de Caccone. Además, el grupo de Sharakhov encuentra que el reordenamiento $2L^a$ es ancestral, contrariamente a la idea que se tenía de que el $2L^+$ fijado en *A. quadriannulatus* sería el que tendría dicha categoría. Tradicionalmente, se ha venido considerando a la especie *Anopheles quadriannulatus* como la especie del Complejo que conservaría las características más ancestrales. Ahora, se sabe que el reordenamiento $2L^+$ de *A. quadriannulatus* es una derivación del $2L^a$, por lo que la posición basal de *quadriannulatus* en la evolución de Complejo queda eliminada.

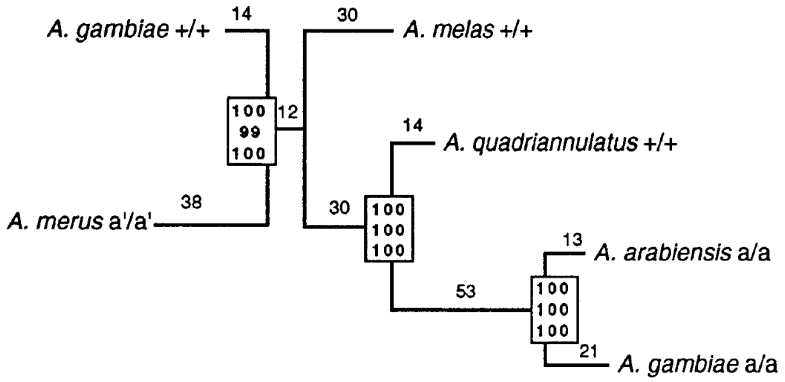



**Figura 8.** Árbol consenso *bootstrap*, obtenido por Caccone et al. (1998), basado en el alineamiento de tres secuencias interiores a la inversión $2L^a$ (secuencias denominadas pkm129, pkm2 y pkm79) y de una secuencia fuera de la inversión pero cercana al punto de rotura (secuencia pkm122). Los símbolos después del nombre de cada una de las especies indica el tipo de reordenamiento. Los valores bootstrap (dentro de los cuadros) representan porcentajes de 1000 réplicas para MP y NJ (arriba y en medio, respectivamente) y de 500 réplicas para ML (abajo). Figura tomada de Caccone et al. (1998).

Ayala y Coluzzi (2005) han reinterpretado recientemente la evolución de las especies que conforman el Complejo de *A. gambiae* (figura 9). Ambos optan por considerar a *Anopheles arabiensis*, descendiente de *Pyretophorus* (procedente de la Península arábiga), como la especie ancestral del Complejo. Según los autores, existen dos evidencias para esta conclusión: la primera es que *A. arabiensis* es el único miembro del Complejo presente en el "cuerno" de África (territorio ocupado por Somalia y la parte más

oriental de Etiopía) y en la Península arábiga; la segunda evidencia es que el reordenamiento $L^a$ del cromosoma *2* está fijado en *A. arabiensis* (Coluzzi et al., 2002), del que se cree que es un reordenamiento ancestral debido a que se encuentra fijado en especies de otros grupos cercanos, como es el Complejo de *Anopheles subpictus*. Los autores argumentan que originariamente *A. arabiensis* era zoofílica y exofílica, y se habría originado en Oriente Medio. Después llegaría a África, a través de la Península arábiga, y más tarde se volvería antropofílica y endofílica tras una adaptación gradual al entorno humano en la región de Sudán y el Oeste de África, en donde actualmente esta especie presenta el mayor polimorfismo cromosómico. *A. arabiensis* se dispersaría inicialmente en el Este África hace más de 6000 años, alcanzando pronto Madagascar, en donde todavía permanece como zoofílica y exofílica, habiendo fallado aquí para adaptarse a los hábitats humanos, tal vez debido a que la densidad baja de humanos no proporcionó la presión selectiva necesaria para dicha adaptación.

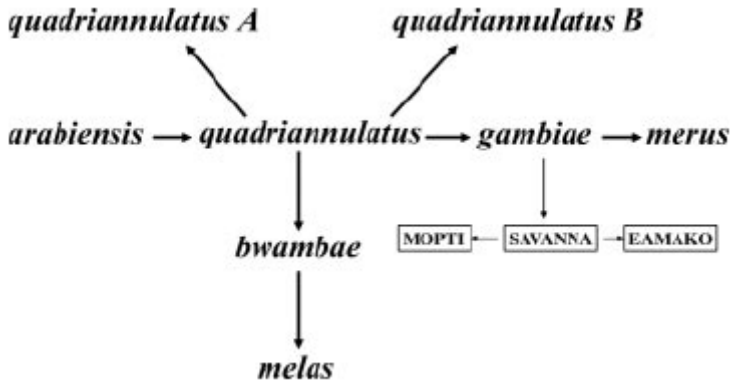



**Figura 9.** Relaciones filogenéticas más probables entre las siete especies del Complejo de *Anopheles gambiae*, deducida por Ayala y Coluzzi (2005). La especie ancestral más probable es *Anopheles arabiensis*, que se diferencia de *Anopheles quadriannulatus* en tres inversiones del cromosoma *X* y se diferencia de *Anopheles gambiae* en otras dos inversiones más en dicho cromosoma. Los factores reproductivos entre estas especies se encuentran principalmente localizados en el cromosoma X. Figura tomada de Ayala y Coluzzi (2005).

Según Ayala y Coluzzi, los patrones de inversión cromosómica muestran que *Anopheles arabiensis* dio lugar a *Anopheles quadriannulatus*, de hecho esta última retiene todavía la condición ancestral de zoofílica y exofílica. Las inversiones asociadas al cromosoma *X*, presentes en *Anopheles quadriannulatus*, contendrían los factores para el

aislamiento reproductivo entre ambas especies. *Anopheles quadriannulatus* daría lugar a dos especies: *quadriannulatus A* en África del Sur, y *quadriannulatus B* en Etiopía; ambas presentan cromosomas homosecuenciales (aunque la especie *A* presenta dos inversiones polimórficas). Estas dos especies alopátricas representan reliquias de la especie ancestral, y llegarían a diferenciarse genéticamente tras su aislamiento geográfico. Además, de la especie *Anopheles quadriannulatus* ancestral se originarían otros dos linajes: uno de ellos llevaría a *Anopheles bwanbae*, especie que presenta una distribución geográfica reducida a el Noreste de Uganda, y a *Anopheles melas*, especie que se distribuye a lo largo de la costa Oeste africana; el otro linaje daría lugar a *Anopheles gambiae* y *Anopheles merus*.

### I.1.2.2. Origen de la especie *Anopheles gambiae.*

El origen de la especie *A. gambiae* ha sido recientemente reinterpretado por Ayala & Coluzzi (2005). Este origen se remonta a no más de 4000 años, en una extensísima masa forestal centroafricana, y está necesariamente ligado a la historia de las poblaciones humanas que deforestaron aquel bosque para dedicarlo a las actividades agrícolas.

La agricultura fue introducida en África hace aproximadamente 8000 años proveniente de Mesopotamia. Sin embargo, el bosque lluvioso centroafricano permaneció impenetrable durante un largo periodo de tiempo, sin existir trazas de la acción de la agricultura hasta hace unos 4000 años (Willis, Gillson & Brncic, 2004). Hace unos 2300 años, dicha masa forestal fue invadida por agricultores del pueblo Bantú, que adoptaron las técnicas de deforestación por tala y quema para dedicar el suelo a la agricultura. La consecuente regresión del bosque, debido a la acción humana, y el aumento de la pluviosidad, fueron factores determinantes para la expansión de la mosca de *tze-tze* (género *Glossina*), vector del protozoo *Trypanosoma brucei*, que diezmó el ganado. Este hecho promovió que la especie *Anopheles quadriannulatus*, inicialmente zoofílica, se adaptara a la alimentación de sangre humana. Así, las nuevas condiciones promovieron una fuerte selección hacia la antropofília y la endofília, favoreciendo la evolución de la especie *A. quadriannulatus* ancestral hacia *Anopheles gambiae* (Willis, Gillson & Brncic, 2004; Ayala & Coluzzi, 2005). Esta interpretación del origen de la especie *A. gambiae* explica la presencia del reordenamiento cromosómico primitivo *2R*, que supone adaptación al bosque lluvioso, a pesar de que la especie *gambiae* solamente puede criar en lugares modificados

por la agricultura humana, dado que sus larvas requieren la luz solar para desarrollarse (Coluzzi et al., 2002).

El proceso de especiación en *A. gambiae* fue mucho más lejos (así como también ha ocurrido con las especies *A. arabiensis* y *A. melas*), y es que en el Sur de Mali y en el norte de Guinea hay tres "formas" de *A. gambiae* cromosómicamente distintas (denominadas *Savanna*, *Mopti* y *Bamako*), que son parcialmente simpátricas o parapátricas y que presentan cierto aislamiento reproductivo que hace suponer una especiación incipiente dentro de la especie, pero estas cuestiones son tratadas más adelante.

### I.1.2.3. Los polimorfismos de inversión cromosómica en la especie *Anopheles gambiae sensu stricto*: las "formas cromosómicas" de *A. gambiae.*

Conocer la estructura poblacional del Complejo de *A. gambiae*, y especialmente de la especie *A. gambiae sensu stricto*, ha sido siempre un objetivo muy importante para poder determinar qué taxones del Complejo son los vectores del *Plasmodium* que causa la Malaria, así como para determinar la importancia relativa de cada uno de ellos en la transmisión del patógeno. Dicho conocimiento permitiría desarrollar unas estrategias de control más adecuadas. El estudio de los reordenamientos cromosómicos en cromosomas politénicos ha tenido gran relevancia en este sentido, sirviendo para la clasificación de *A. gambiae*.

Es bien sabido que las inversiones cromosómicas pueden estar implicadas en los procesos de especiación (Noor et al., 2001; Ayala & Coluzzi, 2005; Kirkpatrick & Barton, 2006; Manoukis et al., 2008). El análisis de los patrones de bandas en cromosomas politénicos ha revelado con frecuencia diferencias fijadas entre especies próximas debidas a inversiones cromosómicas. Desafortunadamente, el estudio de cromosomas politénicos presenta como limitaciones el sexo y/o la etapa del desarrollo y, además, la interpretación del bandeado requiere experiencia y el empleo de un tiempo considerable. Entre los anofelinos, no todos tienen cromosomas politénicos con patrones claros y, lo que es más importante, no todas las especies próximas de anofelinos difieren en dichos patrones (Hunt, Coetzee & Fettene, 1998; Somboon et al., 2001).

La estructura poblacional de *A. gambiae sensu stricto* ha venido siendo estudiada empleando la citogenética clásica para describir polimorfismos de inversión cromosómica. Tanto *A. gambiae* como *A. arabiensis* han desarrollado un patrón muy complejo de polimorfismos de inversión cromosómica, a diferencia de las otras especies que conforman el Complejo de *A. gambiae*. Los estudios sobre los polimorfismos de inversión en *A. gambiae sensu stricto* han permitido obtener una importante conclusión general acerca de la estructura poblacional de la especie: "a escala macrogeográfica pueden diferenciarse las poblaciones del Este de África, con un limitado nivel de polimorfismos de inversión, y las poblaciones del Oeste de África, en donde pueden ser identificados un variado número de reordenamientos cromosómicos" (Coluzzi, Petrarca & DiDeco, 1985; (Petrarca & Beier, 1992). Los estudios citogenéticos en poblaciones del Oeste de África revelaron la existencia de un amplio abanico de cariotipos de inversión, permitiendo definir varias "formas cromosómicas".

En la figura 10 se representa un mapa con la posición para las inversiones fijadas y polimórficas en el cromosoma *2*, aplicable a las especies del Complejo de *A. gambiae*. Las inversiones más frecuentemente observadas en *A. gambiae* son *j*, *b*, *c*, *d* y *u* en *2R* y la inversión *a* en *2L*. Las cinco inversiones en *2R* pueden asociarse en varias combinaciones, excepto *u* y *d* que solapan (*u* está incluida en *d*).


**Figura 10.** Representación esquemática de las principales inversiones fijadas y polimórficas a lo largo del cromosoma *2* de *Anopheles gambiae*. Todas las inversiones son referidas a una ordenación y se designan con letras minúsculas, independientemente para cada uno de los brazos cromosómicos (Coluzzi et al., 1979). Así, cada secuencia cromosómica diferente de la estándar es designada con la letra de la inversión precedida del brazo cromosómico en donde ocurre el reordenamiento (por ejemplo, $2L^{a}$ corresponde a la inversión de la región "a" en el brazo "L" del cromosoma "2"). El símbolo "+" en, por ejemplo, $2L^{+}$, describe el reordenamiento estándar en el brazo "L" del cromosoma "2". Y el símbolo "/" en, por ejemplo, $2L^{a/+}$, significa que la inversión "a" en el brazo "L" del cromosoma "2" es polimórfica. Para designar específicamente la no-inversión, o reordenamiento estándar en, por ejemplo, la región "a", se puede representar como "$+^{a}$".

En el Oeste de África, las poblaciones de *A. gambiae* están compuestas por taxones entre los que existe un aislamiento reproductivo al menos parcial (Coluzzi, Petrarca &

DiDeco, 1985; (Toure et al., 1998; Black & Lanzaro, 2001). Inicialmente, estos taxones se definieron basándose en las combinaciones características de sus inversiones paracéntricas en el cromosoma *2*, para las cuales la mayoría de las frecuencias observadas en las poblaciones diferían significativamente de las frecuencias esperadas asumiendo apareamiento aleatorio. Conforme a esta observación, Coluzzi, Petrarca y Di Deco (1985) definieron cinco formas cromosómicas en *A. gambiae*. Estas nuevas formas cromosómicas recibieron los nombres no linneanos de *Forest*, *Bissau*, *Savanna*, *Bamako* y *Mopti*:

(1) La forma FOREST, caracterizada por los reordenamientos estándar $2R^{+/+}$ $2L^{+/+}$, o por un único polimorfismo de inversión $2R^{b}$, $2R^{d}$ o $2L^{a}$; (2) la forma BISSAU, caracterizada por una elevada frecuencia de la inversión $2R^{d}$ y del reordenamiento estándar $2L^{+}$; (3) la forma SAVANNA, con una elevada frecuencia de las inversiones $2R^{b}$ y $2L^{a}$, y con menor frecuencia se dan polimorfismos que afectan al reordenamiento $2R^{cu}$ y a las inversiones *j*, *d* y la rara inversión *k*; (4) la forma BAMAKO caracterizada por el reordenamiento fijado $2R^{jcu}$ y el reordenamiento $2R^{jbcu}$; (5) la forma MOPTI; que muestra una elevada frecuencia de $2R^{bc}$, $2R^{u}$ y casi fijación del reordenamiento $2L^{a}$.

Tras la definición de las formas cromosómicas de *A. gambiae*, siguió una larga discusión que solo comenzó a vislumbrar su resolución en los primeros años de esta nueva década: ¿son las formas cromosómicas verdaderos taxones aislados reproductivamente?. La existencia o no de aislamiento reproductivo entre las formas cromosómicas, y también el grado de aislamiento si este existiera, supuso un debate que, con el aumento de los conocimientos acerca de la estructura poblacional de *A. gambiae*, llevó a definir las denominadas “formas moleculares” en la especie.

### I.1.2.4. Diferencias en el ADN ribosómico: las "formas moleculares" *M* y *S* de *A. gambiae*. Evidencias moleculares de una incipiente especiación dentro de *Anopheles gambiae sensu stricto*.

La búsqueda de evidencias genéticas sobre la posible existencia de aislamiento reproductivo pre-apareamiento entre las formas cromosómicas de *A. gambiae*, llevó a algunos investigadores a estudiar regiones genómicas no ligadas a las inversiones del cromosoma *2* (della Torre et al., 2001), dado que el hipotético hallazgo de diferencias fijadas, o incluso el de frecuencias significativamente distintas, en *loci* externos a las inversiones entre las formas cromosómicas, también podría soportar la hipótesis del aislamiento reproductivo. Además, el hallazgo de diferencias fijadas entre las diferentes formas cromosómicas, permitiría desarrollar un protocolo que facilitara el diagnóstico específico de las mismas.

Inicialmente, Favia y sus colaboradores (Favia et al., 1997), analizando el ADN ribosomal de *A. gambiae*, observaron unas importantes diferencias moleculares que permitían discernir entre, por un lado, la forma MOPTI y, por otro lado, las formas SAVANNA y BAMAKO. En dicho estudio se analizó un fragmento de 1.3Kb que contenía parte de la región codificadora para *28S* y también parte del *IGS* (*Intergenic Spacer Region*). El fragmento fue amplificado por PCR y cortado con diferentes combinaciones de enzimas de restricción, para detectar polimorfismos en individuos de las formas cromosómicas SAVANNA, MOPTI y BAMAKO (las tres formas presentes en Mali y Burkina Faso). Los patrones obtenidos mostraban diferencias de la forma MOPTI con las otras dos (SAVANNA y BAMAKO). Estas conclusiones permitieron establecer un primer método diagnóstico basado en RFLP-PCR para diferenciar a los individuos de la forma MOPTI de los individuos de las formas cromosómicas BAMAKO y SAVANNA. La validez del método quedó confirmada al ser probado en 203 hembras de mosquitos procedentes de localidades de Mali y Burkina Faso. El descubrimiento de estas diferencias llevó a introducir el término "formas moleculares", para designar a los dos grupos de individuos que el método diagnóstico permitía diferenciar

En el año 2001 se publicaron conjuntamente tres artículos en el número 10 de la revista *Insect Molecular Biology* que, siguiendo la línea de Favia et al (1997), suponían un importante avance en el conocimiento de la estructura poblacional de la especie *Anopheles gambiae* (della Torre et al., 2001; Favia et al., 2001; Gentile et al., 2001).

En el trabajo realizado por Favia et al. (2001), se llevó a cabo un análisis comparativo de secuencias del *IGS*, del ADN ribosomal, pertenecientes a individuos de las tres formas cromosómicas presentes en Mali y Burkina Faso, es decir, las formas MOPTI, SAVANNA y BAMAKO. La región estudiada, que comprendía el extremo 5´ del *IGS*, tenía una longitud de 2.3Kb. El análisis comparativo detallado permitió identificar 10 sustituciones nucleotídicas que diferenciaban a la forma cromosómica MOPTI de las formas SAVANNA y BAMAKO. El grupo de Favia acababa de identificar las primeras diferencias fijadas dentro de *A. gambiae sensu stricto*. Teniendo en cuenta estas diferencias, se diseñó un protocolo basado en la PCR que permitía, de una manera sencilla, diferenciar entre los mosquitos pertenecientes a las formas cromosómicas MOPTI y [SAVANNA + BAMAKO].

Por su parte, el grupo de della Torre, para aclarar estos patrones de especiación incipiente dentro de *A. gambiae*, extendió la metodología descrita por Favia a mosquitos tomados de países del Oeste y del Este de África (della Torre et al., 2001). Estudiaron un total de 1162 individuos de la especie *A. gambiae sensu stricto*, tomados en un total de 20 países, y no detectaron un patrón mixto (que sería indicativo de la existencia de híbridos) para ninguno de ellos. Esto llevó a que los términos *M* y *S* fueran propuestos para indicar los dos tipos de *IGS*. Además, observaron que, fuera de Mali y Burkina Faso, las tipos *M* y *S* no siempre se correspondían con las formas cromosómicas MOPTI y [BAMAKO + SAVANA]. Este trabajo de della Torre y sus colaboradores asentó definitivamente los términos de "forma molecular *M*" y "forma molecular *S*", para designar a las dos especies incipientes que parecen diferenciarse dentro de la especie *Anopheles gambiae*, términos que en parte ya habían sido ambiguamente introducidos por Favia et al. (1997), cuando designaban respectivamente como *MM* y *SS* a los patrones de RFLP-PCR obtenidos para las formas cromosómicas MOPTI y [BAMAKO + SAVANA] (la nomenclatura *SM* había sido reservada en ese trabajo para los posibles híbridos MOPTI/[BAMAKO + SAVANNA], híbridos que finalmente no fueran detectados).

El intento por encontrar evidencias de aislamiento genético entre las formas cromosómicas de *A. gambiae*, llevó también al grupo de Gentile a secuenciar varias regiones genómicas (fuera del cromosoma *2*) y mitocondriales en individuos de las tres formas cromosómicas MOPTI, SAVANNA y BAMAKO (Gentile et al., 2001). De las regiones analizadas, solo una de ellas presentaba diferencias nucleotídicas que distinguía a

las formas cromosómicas, se trataba del *Internal Transcribed Spacer* (ITS) del ADN ribosomal: tres puntos dentro de dicha secuencia permitían diferenciar a la forma cromosómica MOPTI de las formas SAVANNA y BAMAKO. La observación de estas diferencias, y de las también identificadas por Favia et al. (2001), llevo al grupo de Gentile a definir dos grandes Tipos en el ADN ribosomal, que denominaron Tipo *I* y Tipo *II*. Estos dos Tipos también se corresponden con las formas moleculares *M* y *S* definidas por della Torre et al. (2001). Al igual que el grupo de della Torre, Gentile y sus colaboradores observaron que la correspondencia entre la forma cromosómica y el Tipo de ADN ribosomal solamente se mantenía en Mali y Burkina Faso, de manera que fuera de estos países dicha relación no siempre se cumplía (concretamente este hecho se observó en Tanzania y Madagascar). Finalmente, los autores concluyen que, dado que no han encontrado heterocigotos portando ambas formas, *ITS I* e *ITS II*, todo parecía indicar que estaba claro que *A. gambiae* no se comporta como una unidad panmíctica.

Con el tiempo, se ha venido confirmando que la correspondencia entre las formas moleculares y las formas cromosómicas, tal y como había sido determinada por Favia et al. (1997), no siempre coinciden (della Torre et al., 2002). Es cierto que, dentro de Mali y Burkina Faso las formas moleculares *M* y *S* mantienen la correspondencia con las formas cromosómicas MOPTI y [SAVANNA + BAMAKO], respectivamente, sin embargo, fuera de esas regiones no existe tal correspondencia (figura 11).

El interés por revelar los niveles de flujo génico entre ambas formas moleculares y, por tanto, el estado del proceso de especiación que parece estar teniendo lugar, llevó a Tripet et al. (2001) a estudiar genéticamente 251 hembras de *A. gambiae*, y también el esperma contenido en sus respectivas espermatecas. Las hembras de mosquito estudiadas procedían de Mali, de una localización en donde convivían conjuntamente individuos de ambas formas moleculares. Los autores encontraron que un 1.03% de las hembras de la subpoblación *M* (2/195) contenían esperma de la forma *S*. Igualmente, encontraron un 1.82% de hembras de la subpoblación *S* con esperma de la forma *M*. Pero el hallazgo más interesante fue el de una hembra adulta híbrida *M*/*S*. Los autores concluyeron que la escasa divergencia presente actualmente entre las formas *M* y *S* puede ser explicada por el mantenimiento del flujo génico entre ambas.

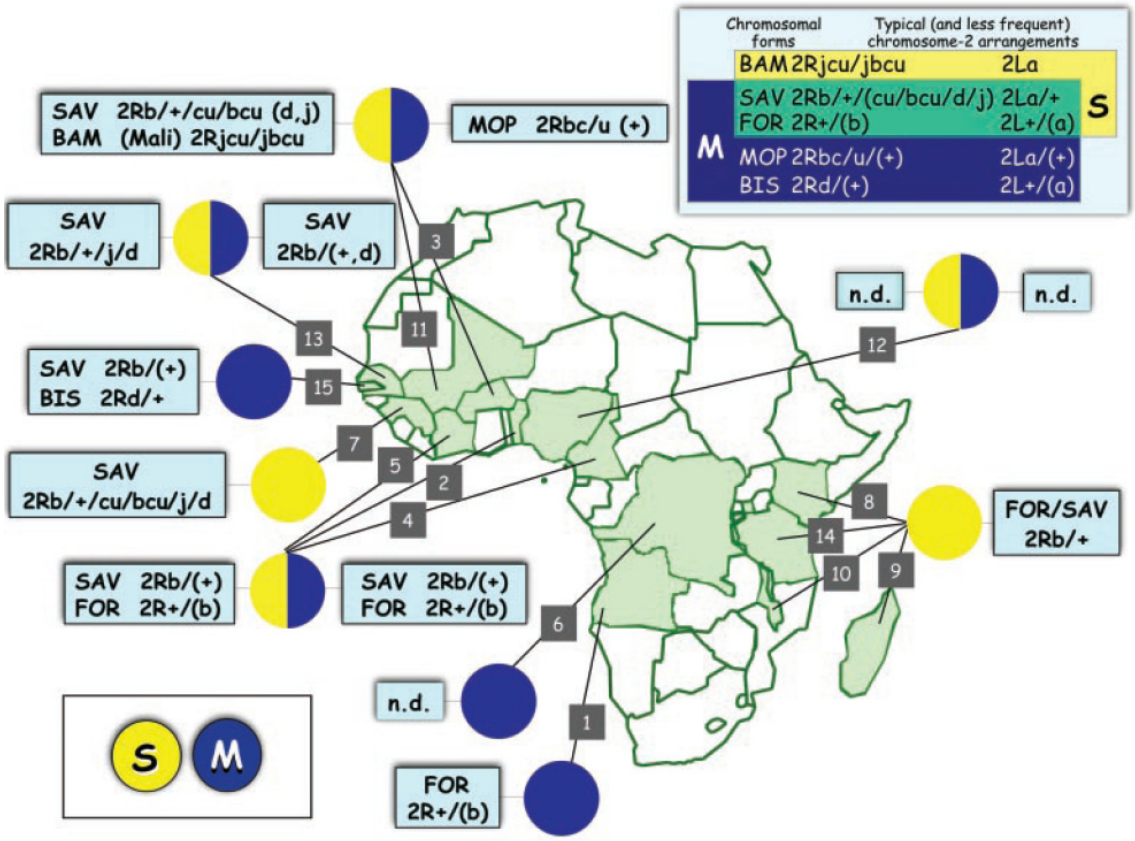



**Figura 11.** Distribución geográfica de las formas moleculares *M* y *S* de *Anopheles gambiae sensu stricto* y su relación con las formas cromosómicas. FOR: Forest; SAV: Savanna; MOP: Mopti; BAM: Bamako; Bis: Bissau; n. d.: cariotipo no determinado. Tomada de della Torre et al. (2002).

Wondji, Simard & Fontenille (2002) estudiaron la variabilidad genética en 10 loci microsatélites distribuidos a lo largo del genoma de *A. gambiae*. Los mosquitos estudiados fueron capturados en cuatro poblados en Camerún que estaban separados por distancias de 35-350 kilómetros. Obtuvieron niveles de diferenciación genética estadísticamente significativos ($F_{st}$>0.035) en las comparaciones entre mosquitos de diferentes formas moleculares dentro de un mismo poblado. La diferenciación genética resultó ser menor ($F_{st}$<0.017) en las comparaciones entre moquitos de una misma forma molecular entre distintos poblados.

El análisis de secuencias genómicas de evolución rápida, como son el ADN ribosomal y los microsatélites, y la constatación de que prácticamente no existen híbridos *M*/*S* en la naturaleza, estaban revelando la existencia de diferenciación entre ambas formas cromosómicas. Sin embargo, el análisis extendido a otros genes del genoma no mostraba

diferencias consistentes entre ambas formas moleculares (Gentile et al., 2001; Mukabayire et al., 2001). Solamente el alelo *Kdr* del gen *voltaged-gated sodium chanel* parecía ser la excepción a la regla.

El gen *voltaged-gated sodium chanel* se encuentra localizado en el brazo *L* del cromosoma *2*. El alelo *Kdr* (*Knock-down resistance*) confiere resistencia a determinados insecticidas. El estudio poblacional llevado a cabo por della Torre et al. (2001) sobre las frecuencias de dicho alelo en Costa de Marfil y en Benin, permitió confirmar la ausencia del alelo *Kdr* en la forma molecular *M* en Costa de Marfil, mientras que en Benín ambas formas moleculares *M* y *S* presentan dicho alelo. La ausencia del alelo *Kdr* en los individuos de la forma molecular *M* en Costa de Marfil, es una evidencia de la restricción al flujo génico existente entre ambas formas moleculares. Por otro lado, la presencia del alelo *Kdr* en individuos de la forma cromosómica *M* en Benin podría ser explicada por introgresión desde la forma *S*.

La importancia de las inversiones cromosómicas en la adaptación ecológica dentro del Complejo de *Anopheles gambiae* ha sido bien establecida (Powell et al., 1999), sugiriendo que las diferentes formas cromosómicas son indicativo de la adaptación a diferentes hábitats. Sin embargo, las formas moleculares *M* y *S* serían el reflejo de la existencia de barreras al flujo génico, indicativo de una especiación incipiente (della Torre et al., 2001).

Aunque es cierto que del cruzamiento entre individuos *M* y *S* resulta una progenie fértil, los híbridos *M*/*S* son raramente observados en la naturaleza. En aquellos lugares en donde los mosquitos de las formas moleculares *M* y *S* viven en simpatría, se ha estimado que la tasa de inseminación heterogamética es de aproximadamente el 1% (Tripet et al., 2001), lo que demuestra claramente la existencia de una barrera pre-apareamiento entre ambas formas, aunque ésta pueda ser incompleta. Todas las evidencias hasta ahora comentadas parecen estar de acuerdo con que ambas formas moleculares se encuentran en las primeras etapas de sus respectivos procesos de especiación, compartiendo un polimorfismo ancestral, debido a que ambas formas proceden de un cercano antepasado común, y manteniéndose un pequeño nivel de flujo génico que continúa homogenizando regiones del genoma que no están implicadas directamente en el proceso de especiación. Esto explicaría porque el estudio de regiones genómicas escogidas al azar no muestra

diferenciación, en contraste con el alelo *Kdr* y las secuencias de evolución rápida en el ADN ribosomal y en secuencias microsatélite.

En este sentido resulta muy clarificador un trabajo recientemente publicado (della Torre, Tu & Petrarca, 2005) que resume todos los datos conocidos hasta el momento (algunos nunca antes reportados) acerca de la distribución geográfica, a nivel macrogeográfico y también local, de las formas moleculares *M* y *S* de *A. gambiae*, así como de los híbridos *M*/*S* identificados. Entre los casi 8000 individuos procedentes de los 9 países pertenecientes al Noroeste africano que han sido estudiados, solamente se han identificado 6 híbridos *M*/*S*, mientras que no han sido identificados híbridos en ninguno de los 7 países del centro-oeste africano (10522 individuos analizados) ni a los 9 países del Este (1060 individuos). De los seis híbridos identificados, tres fueron capturados en Mali, Burkina Faso y Benin, en donde se sabe que *M* y *S* son simpátricas. Los otros tres híbridos fueron encontrados en lugares de Gambia, Guinea y Costa de Marfil en donde solamente se ha identificado una de las dos formas moleculares, aunque la posibilidad de la existencia de la otra forma en zonas cercanas no puede ser excluida. La frecuencia de híbridos *M*/*S* en poblaciones naturales resulta ser inferior al 1% en las poblaciones del Oeste de África, siendo esta una frecuencia comparable a las obtenidas para los híbridos de otras especies del complejo *A. gambiae* claramente distintas, por ejemplo, los híbridos *A. gambiae*/*A. arabiensis* (Gillies, 1987; Petrarca et al., 1991; Powell et al., 1999) y *A. gambiae*/*A. melas* (Bryan et al., 1987).

#### I.1.2.4.1. Distribución de las formas moleculares de *Anopheles gambiae sensu stricto*.

Los datos resumidos por della Torre et al. (2005) muestran, de una manera muy clara (figura 12), que las formas moleculares *M* y *S* tienen diferente distribución a escala macrogeográfica y también local. Los autores recomiendan tener en todo momento en cuenta que la ausencia de una de las dos formas en las muestras analizadas en cada punto no implica que, definitivamente, esa forma no se encuentre presente en la zona. Del total de las muestras extraídas se puede concluir que la forma *M* muestra la distribución latitudinal más amplia (desde 16º Norte a 16º Sur), siendo la única forma encontrada en las sabanas secas en el Norte de Senegal y en áreas limítrofes al desierto tropical en Angola. La forma *S*, sin embargo, presenta la mayor distribución longitudinal (desde 13º Oeste a

50º Este), siendo la única forma observada al Este del Valle del Rift (Weill et al., 2000; della Torre et al., 2001; Lehmann et al., 2003; Leong et al., 2003), con la única excepción de unos pocos individuos encontrados en Zimbabue (Masendu et al., 2004).

En el Oeste africano ambas formas, *M* y *S*, presentan con frecuencia una distribución simpátrica desde el Sur de las sabanas de Sudán y Guinea, localizadas en los 13º Norte, hasta las zonas más húmedas y forestes localizadas por encima de los 4º Sur. Aunque también es cierto que, a lo largo de ese mismo rango de distribución (13º Norte-4º Sur), en algunos puntos solo se ha encontrado una de ambas formas moleculares. A escala local resulta curioso encontrar que en varias áreas desde Mali a Camerún se presentan frecuencias relativas invertidas para ambas formas moleculares, incluso si se comparan localidades vecinas separadas por menos de 50 kilómetros. En Mali y Burkina Faso, estas diferencias han sido atribuidas a diferentes condiciones necesarias para el desarrollo de las larvas, principalmente la disponibilidad de agua en relación con la actividad humana: es decir, parece ser que la forma molecular *M* está asociada a una mayor permanencia de agua que se da en, por ejemplo, zonas de cultivo de arroz o lagos artificiales; mientras que la forma *S* podría desarrollarse mejor que la *M* en lugares en donde la disponibilidad de agua fuera dependiente de la pluviometría. Además, en otros países africanos, algunos autores reportaron una mayor proporción de la forma *M* en lugares de cría más contaminados (Kristan et al., 2003), o en localizaciones más urbanizadas (Wondji et al., 2005).

No se han realizado estudios a gran escala acerca de los cambios en las frecuencias relativas de ambas formas moleculares en función de los cambios estacionales de las condiciones ambientales. Sin embargo, en Mali se ha encontrado un incremento de frecuencia relativa de la Forma *M* en la estación seca. Este hecho parece estar relacionado con la mayor dependencia de la forma *S* por la lluvia para el desarrollo de sus larvas (Toure et al., 1998).

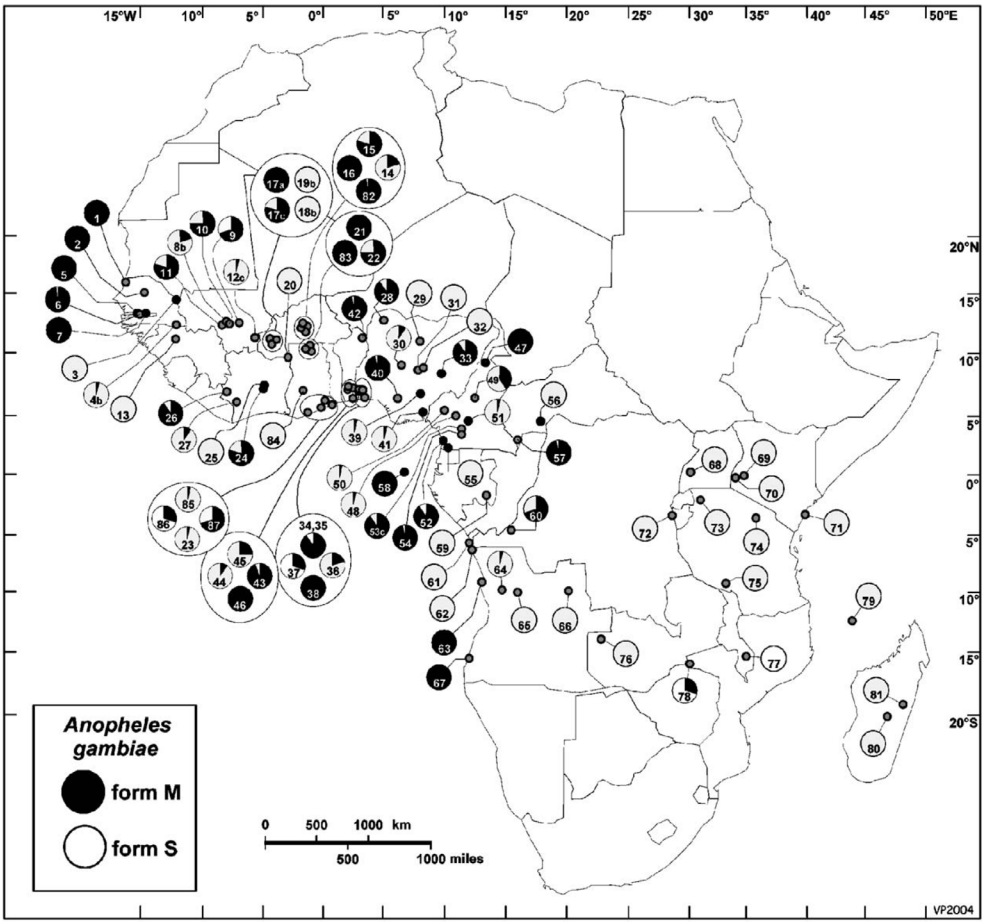



**Figura 12.** Frecuencias relativas de las formas moleculares *M* y *S* de *Anopheles gambiae sensu stricto* en 87 localizaciones a lo largo de 24 países de África. Figura tomada de della Torre et al. (2005).

### I.1.3. Estructura de las poblaciones de *A. gambiae*: flujo génico y tamaño efectivo de población.

La estructura poblacional de la especie *Anopheles gambiae* es muy compleja. El nivel taxonómico más alto del sistema es *Anopheles gambiae sensu lato*, que comprende al menos siete especies morfológicamente indistinguibles y entre las cuales existe cierto grado de flujo génico. Una de estas especies es *Anopheles gambiae sensu estricto*, que presenta hasta cinco formas cromosómicas con diferentes niveles de tolerancia a distintas condiciones ecológicas. También existe un proceso de especiación incipiente dentro de la especie, distinguiéndose las formas moleculares *M* y *S* de *A. gambiae*, que presentan restricción de flujo génico entre ambas. Además, los cambios estacionales anuales propios

del clima, los largos períodos de sequía y la presencia de infranqueables barreras físicas, como la representada por el impenetrable Valle del Rift, son factores que determinan la estructuración poblacional de esta especie en África.

### I.1.3.1. Patrones de flujo génico a escala macrogeográfica.

Un temprano trabajo de Coluzzi, Petrarca & Di Deco (1985) mostraba que las poblaciones de *A. gambiae*, a lo largo de grandes distancias, difieren extraordinariamente en cuanto a la distribución de determinadas inversiones cromosómicas paracéntricas. Sin embargo, una década más tarde, un trabajo de Lehmann et al. (1996) concluía un escenario diferente: estudiando variación a nivel bioquímico (isozimas) y genético (ADN mitocondrial y microsatélites) entre poblaciones de Kenia y Senegal, que estaban separadas por más de seis mil kilómetros, obtuvieron una escasa diferenciación a lo largo de las poblaciones analizadas ($F_{ST}$ = 0.016)[f] y estimaron un elevado nivel de flujo génico entre ellas [g]($N_m$ > 7.7). Más tarde, Besansky et al. (1997) analizaron un fragmento de 665pb perteneciente al gen mitocondrial *ND5*, comparando individuos de siete localidades de Kenia y tres de Senegal, encontrando también homogeneidad entre poblaciones de Kenia y Senegal ($F_{ST}$ = 0.085) y, consecuentemente, una alta estima del flujo génico entre dichas poblaciones ($N_m$ = 5.4).

Así, los estudios comparativos a gran escala no encontraron evidencias de aislamiento debido a distancia, es decir, no observaron relación entre niveles de divergencia genética ($F_{ST}$) y distancia geográfica. Los resultados de estos trabajos sugerían que *A. gambiae*, a lo largo de su rango de distribución, estaría compuesta de poblaciones que intercambian entre sí individuos a una tasa suficiente como para evitar que diverjan genéticamente. Los autores sugerían que la ausencia de divergencia entre poblaciones muy separadas podría ser consecuencia de una expansión reciente de la especie, hace unos 2000

[f] $F_{st}$ es una medida de la estructuración genética de las poblaciones, desarrollada por Sewall Wright (1969,1978). $F_{st}$ es la proporción de la varianza genética total contenida en una subpoblación, con respecto a la varianza genética total. Toma valores de 0 a 1, de manera que valores elevados de $F_{st}$ implican un grado elevado de diferenciación entre poblaciones.

[g] *Nm* es el producto de *N* (Tamaño efectivo de población) y *m* (proporción efectiva de inmigración). Si *Nm*>1, entonces hay suficiente flujo génico como para vencer los efectos de la deriva genética, y si *Nm*>4, entonces las poblaciones locales pertenecen a una única población panmíctica (Wright, 1931). *Nm* se estima como: $Nm=(1/F_{st}-1)/4$ en diploides, y como $Nm=(1/F_{st}-1)/2$ en haploides (y para ADN mitocondrial).

a 5000 años, asociada a la expansión de poblaciones humanas en África (Coluzzi, Petrarca & Di Deco, 1985). Como alternativa a esta hipótesis, los autores proponían que el flujo génico entre poblaciones a través del continente africano podría ser contemporáneo, como resultado de una migración activa y un transporte pasivo asociado a la actividad humana, apoyándose en el conocimiento de la reciente introducción de *A. gambiae* en Brasil como consecuencia de la actividad humana en la década de 1930.

A pesar del alto grado de interés de estos estudios, las hipótesis propuestas no eran definitivas. La conclusión sobre la existencia de un flujo génico contemporáneo, a través de vastas extensiones, no era consistente con el conocimiento que se tiene acerca de la capacidad de los individuos de *A. gambiae* para dispersarse, ya que el desplazamiento más largo obtenido mediante observación directa fue estimado en el rango de 3.6 a 7 kilómetros (Gillies, 1961; Toure et al., 1998), aunque también algunas estimas indirectas sugieren capacidades mayores, del orden de decenas o como mucho cientos de kilómetros (McLain et al., 1989; Carnahan et al., 2002). Además, una reflexión más profunda sobre la metodología en la que se apoyan los trabajos que concluyen un extensivo flujo génico entre poblaciones muy distantes de *A. gambiae*, permite plantear una serie de dudas en torno a tal conclusión:

Por ejemplo, en lo referido a las estrategias para la recogida de muestras, los dos trabajos comentados anteriormente (Lehmann et al., 1996; Besansky et al., 1997) incluyen en sus estudios solamente dos regiones (Kenia y Senegal) que coinciden con los extremos del rango de distribución de *A. gambiae*. Sin embargo, un estudio más amplio, que incluya más poblaciones a lo largo de toda la distribución de la especie, podría revelar un patrón que llevara a pensar en una explicación diferente para la aparente ausencia de divergencia. Así, por ejemplo, los resultados obtenidos por Besansky et al. (1997) no concordaban con otros revelados por Donnelly & Townson (2000).

En su estudio, Donnelly y Townson incluyeron nueve localidades del Este de África a lo largo de unos 4500 kilómetros, desde Sudán hasta Mozambique, estudiando 8 loci microsatélites y tamaños de muestra mayores (de 29 a 59 individuos por población y por locus), encontrando grandes diferencias significativas en las frecuencias genotípicas entre poblaciones separadas por 200 kilómetros y, además, en Mozambique incluso en poblaciones separadas por tan solo 25 kilómetros. También obtuvieron una correlación positiva significativa entre los valores de $F_{ST}$ y la distancia geográfica.

El estudio de Besansky et al. (1997) sale todavía peor parado viendo que el número de mosquitos estudiado es demasiado bajo. Ya que, para algunas localidades, tan solo se analizaron 4 mosquitos, y para ninguna de las localidades se analizaron más de 10. Años más tarde, los autores reanalizaron las mismas localidades en Kenia incluyendo un tamaño de muestra mayor (Lehmann et al., 2000). Cuando se aumentó el tamaño de la muestra (de un total de 37 a 71) la estima para la divergencia resultó mucho mayor y, por consiguiente, la estima para el flujo génico resultó menor. Los autores concluyeron que la discrepancia entre ambos trabajos se debía al efecto del pequeño tamaño de muestra empleado en el estudio original.

El trabajo más ambicioso a gran escala geográfica sobre *A. gambiae* fue publicado por (Lehmann et al., 2003). En este trabajo se estudiaron 11 loci microsatélites, en mosquitos tomados en 16 localizaciones diferentes de 10 países africanos. Los resultados de divergencia revelaron una subdivisión de la especie a lo largo del continente africano. Por un lado, identificaron un grupo poblacional Noroccidental, que comprendía las poblaciones de Senegal, Gana, Nigeria, Camerún, Gabón, República Democrática del Congo y el Oeste de Kenia. Por otro lado, un grupo poblacional Suroriental que incluía a las poblaciones del Este de Kenia, Tanzania, Malawi y Zambia. La diferenciación entre estos dos grupos poblacionales resultó alta ($F_{ST} > 0.1$). La diferenciación genética entre las poblaciones dentro de cada uno de los dos grupos fue substancialmente menor, y se observó una relación significativa entre distancia genética y distancia geográfica. Los autores sugirieron que la diferenciación entre el grupo Noroccidental y el Suroriental sería consecuencia de un cuello de botella reciente en el grupo Suroriental y de la existencia de barreras físicas que limitan el flujo génico entre ambos grupos. Además, parecía que la población de Zambia ocupaba una posición intermedia entre ambos grupos, representando posiblemente un puente para el flujo génico entre ambos grupos.

En resumen, los estudios llevados a cabo con el propósito de describir la estructura genética de *A. gambiae* a escala macrogeográfica han proporcionado resultados enfrentados. El trabajo inicial de Coluzzi, Petrarca & Di Deco (1985) basado en la distribución de las inversiones cromosómicas sugería unas grandes diferencias entre las poblaciones del Este de África y del Oeste de África. Los trabajos posteriores basados en el estudio de isozimas, microsatélites y ADN mitocondrial (Lehmann et al., 1996; Besansky et al., 1997) llevaban a pensar que *A. gambiae* existía como una única gran población, más

o menos diferenciada genéticamente, a lo largo de todo su rango de distribución. Además, un trabajo más reciente (Lehmann et al., 2003), donde se estudiaban frecuencias microsatélites entre 16 poblaciones distribuidas a lo largo de todo el continente, reveló la existencia de dos grandes grupos de poblaciones a escala macrogeográfica. Es este último trabajo, el que parece describir con mayor precisión la realidad sobre la estructura genética de *A. gambiae* a lo largo del continente africano. Es más: en los primeros trabajos, que comparaban dos poblaciones separadas por 6000 kilómetros, la población de Kenia estudiada era Asembo, localizada al Oeste del Valle del Rift. Hoy se sabe que el Valle del Rift actúa como barrera al flujo génico entre las poblaciones de *A. gambiae* separadas por el mismo. Así, si estos primeros trabajos incluyeran muestras de poblaciones cercanas a la costa de Kenia, es decir, al otro lado del Valle del Rift, el resultado sería bien diferente.

A continuación, se describen algunos trabajos acerca de la diferenciación genética de las poblaciones de *A. gambiae* a escala macrogeográfica, tratando separadamente las dos grandes regiones de interés en el continente africano. En primer lugar, se tratará el Este de África y la importancia del Valle del Rift como barrera al flujo génico. A continuación, se tratarán las poblaciones del Oeste y Centro de África, con una estructura más compleja que las poblaciones del Este de África.

#### I.1.3.1.1. África oriental y el Valle del Rift.

McLain et al. (1989) estudiaron por *RFLPs* la región *IGS* del ADN ribosomal de diversas poblaciones de *A. gambiae* en Kenia. El estudio incluía poblaciones de 7 localidades del Oeste de Kenia y 8 de la costa de Kenia. Encontraron que las poblaciones separadas por 10 ó más kilómetros diferían significativamente en las frecuencias de los *RFLPs* y que las poblaciones del Oeste de Kenia, localizadas a 700 kilómetros de las demás, no compartían *RFLPs* con las de la costa de Kenia. Los autores concluían que las poblaciones de *A. gambiae* en dicha región seguían un modelo de aislamiento por distancia con flujo génico restringido entre poblaciones separadas por tan solo 10 kilómetros y que las poblaciones separadas por unos pocos cientos de kilómetros se encontrarían completamente aisladas genéticamente.

Lehmann et al. (1997) llevaron a cabo un estudio poblacional en la misma parte del Oeste de Kenia. Estudiaron la estructura poblacional basada en 5 loci microsatélites y la

variación en una secuencia de 648pb perteneciente al locus mitocondrial *ND5*. Sus análisis iban dirigidos a obtener valores de divergencia genética a varios niveles: entre casas de una misma localidad, entre localidades y entre localidades que distaban a más de 50 kilómetros. Evidentemente, no encontraron divergencia genética entre individuos recogidos en diferentes casas dentro de una misma localidad, pero sus resultados a una escala mayor contradecían los obtenidos por McLain et al. (1989). Los resultados sugerían que no existían diferencias significativas entre localidades a distancias mayores de 50 kilómetros, llevando a los autores a concluir que el flujo génico era extensivo entre poblaciones separadas a dichas distancias. Estos resultados concordaban, además, con el estudio previo donde no encontraran diferencias significativas entre poblaciones separadas por 6000 kilómetros. Así, al igual que en su anterior trabajo, Lehmann y sus colaboradores argumentaron que estas observaciones podrían ser explicadas por una alta tasa de flujo génico contemporáneo y relativamente raros episodios de extinción-recolonización.

Kamau et al. (1998) llevaron a cabo un estudio similar al de Lehmann de 1997, en la misma región de Kenia. Esta vez, analizaron 7 loci microsatélites en poblaciones de 7 localidades de Kenia separadas por menos de 10 kilómetros en el área de Asembo. Los autores no detectaron evidencias de divergencia entre poblaciones en dichos lugares ($F_{ST}$ = 0.0016, $N_m$ = 5.66). Sin embargo, en este estudio también se incluyeron individuos procedentes de Kilifi, una localidad que se encuentra 700 kilómetros al Este de Asembo, en la costa del Océano índico. Al igual que en el estudio de McLain et al. (1989), los autores detectaron diferencias significativas entre las poblaciones del Este y del Oeste ($F_{ST}$ = 0.075, $N_m$ = 1.54). Estos resultados sugerían que el flujo génico entre las poblaciones del Este y del Oeste de Kenia está severamente restringido y genera muchas dudas sobre los primeros trabajos que concluían un flujo génico extensivo entre las poblaciones de *A. gambiae* a lo largo del continente africano. Además, Kamau et al. (1998) proporcionaron por primera vez pruebas de que el Gran Valle del Rift, que separa el Este y el Oeste de Kenia, estaría actuando como barrera al flujo génico entre las poblaciones de *A. gambiae* distribuidas a ambos lados del mismo.

Este fenómeno fue tratado por Lehmann et al. (1999) en un trabajo mucho más detallado. Estudiaron seis poblaciones de *A. gambiae*, cuatro al Oeste del Valle del Rift y dos al este. Se obtuvieron las frecuencias de nueve loci microsatélites para las comparaciones entre dichas poblaciones. El objetivo del trabajo era determinar si el alto

grado de diferenciación genética entre poblaciones a ambos lados del Valle del Rift se debe a que éste actúa como barrera al flujo génico o si se debe a otros factores, como diferencias en el tamaño efectivo de población ($N_e$) entre las poblaciones estudiadas o como, simplemente, la distancia que las separa. Aunque los autores detectaron cierta diferencia en el $N_e$ entre poblaciones a ambos lados del Valle del Rift, estas diferencias por sí solas no podrían explicar el nivel de divergencia entre poblaciones a ambos lados del Valle del Rift, concluyendo que el Gran Valle del Rift representa, efectivamente, una barrera al flujo génico, y que este fenómeno explica la divergencia genética observada entre las poblaciones de *A. gambiae* del Oeste y el Este de Kenia. Estos resultados son consistentes con los trabajos previos de McLain et al. (1989) y de Kamau et al. (1998), pero entran en conflicto con el trabajo de Besansky et al. (1997), que no encontraba diferencias significativas entre las poblaciones a ambos lados del Valle del Rift. Esta incongruencia fue resuelta, tal y como se ha explicado, cuando se repitió el estudio incrementando el tamaño de muestra (Lehmann et al., 2000).

Tomando todos estos estudios en conjunto se revela como evidencia sólida el funcionamiento del Gran Valle del Rift como una importante barrera frente al flujo génico entre las poblaciones de *A. gambiae* que separa. A la luz del trabajo de Lehmann et al. (2003), en donde describen la estructura genética de *A. gambiae* a lo largo de una gran parte de su rango distribución (ver en párrafos anteriores), parece que el Gran Valle del Rift, que se extiende desde el oeste de Kenia hasta Mozambique, es una importante barrera al flujo génico, conformando una división Oeste/Este de las poblaciones de *A. gambiae*. El dato obtenido por Lehmann et al. (2003) para el Oeste de Zambia, al Suroeste del Valle del Rift, sugiere que dicha región podría representar un puente al flujo génico entre ambas poblaciones del Oeste y del Este de la barrera, es decir, por un lado las poblaciones del Este del continente africano y, por otro, las poblaciones del Oeste y el Centro del continente.

#### I.1.3.1.2. Oeste y Centro de África.

Los estudios iniciales sobre los polimorfismos de inversión cromosómica en Nigeria revelaron la existencia de dos formas cromosómicas, las formas SAVANNA y FOREST (Coluzzi et al., 1979; Coluzzi, Petrarca & DiDeco, 1985). La abundancia relativa

de ambas formas es clinal, con predominio de la forma SAVANNA en el área Norte y más seca de Nigeria, decreciendo gradualmente su proporción frente a la forma FOREST hacia el Sur, donde predomina esta última. Onyabe & Conn (2001) intentaron determinar el nivel de flujo génico entre poblaciones en Nigeria. Analizaron 10 loci microsatélites en individuos pertenecientes a 8 localizaciones diferentes a lo largo de un transecto de 833 kilómetros. El tamaño de muestra en cada punto estaba en el rango de 39-46 mosquitos. La mitad de los loci estudiados estaban localizados dentro de inversiones y la otra mitad fuera de las mismas. Los autores obtuvieron una correlación significativamente elevada entre $F_{ST}$ y distancia geográfica, sugiriendo que la estructura de las poblaciones se ajusta a un modelo de aislamiento por distancia. Sin embargo, un examen más detallado reveló que los valores elevados de $F_{ST}$ eran debidos a tres loci localizados dentro de inversiones y, al eliminar del análisis estadístico dichos *loci*, la correlación entre $F_{ST}$ y distancia geográfica ya no era significativa. Los autores argumentaron que el elevado valor de la divergencia para los *loci* asociados con inversiones era consecuencia del *hitchhiking* de genes localizados dentro de inversiones. Finalmente, Onyabe & Conn concluían que el flujo génico entre las poblaciones de *A. gambiae* en Nigeria no parece estar limitado por la distancia geográfica. Sin embargo, una limitación de este estudio fue el de no cariotipar las muestras estudiadas. Es posible que algunas de las muestras consistieran en mezcla de individuos de la forma SAVANNA y FOREST. De hecho, Coluzzi, Petrarca & Di Deco (1985) encontraron evidencias de que en algunos lugares de Nigeria las formas SAVANNA y FOREST existen en simpatría, estando este hecho asociado con un déficit de heterocariotipos. Este efecto podría estar ocultando cualquier posible correlación entre $F_{ST}$ y distancia geográfica.

Carnahan et al. (2002) llevaron a cabo un estudio en Mali similar al de Onyabe & Conn (2001). Incluyeron individuos de 11 poblaciones procedentes de 6 localidades en un transecto de 536 kilómetros a lo largo del valle del Río Níger. El hábitat a lo largo del transecto era uniforme y no se encontraron evidencias de barreras obvias para el flujo génico dentro del área estudiada. En el estudio se analizaron entre 5 y 23 *loci* microsatélites, el tamaño de muestra era variable entre 4 y 190 individuos. Los puntos de muestreo con menor tamaño de muestra solamente fueron incluidos si el número de loci analizados >20. Encontraron una correlación significativa entre $F_{ST}$ y distancia geográfica para el conjunto de todos los loci y para el conjunto de los loci del cromosoma *3* y del

cromosoma *X*, cuando el análisis se realizó para los microsatélites en cada cromosoma separadamente. Los valores de $N_m$ resultaron en un rango entre 64.43 a 1.26 entre cada par de poblaciones a lo largo del transecto. Los autores concluyen que en esta parte de África la diferenciación genética es consistente con un modelo de aislamiento por distancia. Sin embargo, existen algunas cuestiones no demasiado claras en este trabajo. Así, por ejemplo, no todos los loci analizados fueron los mismos para cada población estudiada, lo que sin duda podría afectar al valor de la $F_{ST}$. Además, el tamaño de muestra en algunos puntos de muestreo era tan bajo como de 4 individuos.

### I.1.3.2. Patrones de flujo génico a escala local.

El patrón de flujo génico a escala local entre poblaciones de *A. gambiae sensu stricto* es, principalmente, reflejo de la existencia de un proceso de especiación dentro de la especie. El conocimiento de este proceso llevó a la definición de las formas moleculares *M* y *S* de *A. gambiae*. Esta diferenciación ocurre solamente en el Oeste de África, aunque en algunas partes del Oeste africano aparentemente solo exista una de ambas formas. En el Este de África, no existen evidencias de diferenciación genética dentro de sus poblaciones, presentándose solamente una de las dos formas moleculares: la forma *S*. La frecuencia de híbridos entre ambas formas moleculares fue estimado en 0.05-0.3%, dependiendo de la población estudiada, una tasa suficiente como para explicar la carencia general de diferenciación genética entre ambas formas, reportada en numerosos estudios.

### I.1.3.3. Tamaño efectivo de población de *A. gambiae*.

El tamaño efectivo de población ($N_e$) resulta un parámetro de mucho interés en el estudio de las poblaciones de todas las especies y, en especial, de *A. gambiae*, dado que el tamaño de población de esta especie se ve afectado con el cambio estacional del clima. Durante la estación seca, la densidad de población de *A. gambiae* es muy baja. En algunos lugares incluso no se encuentran individuos durante dicho período. Tras las primeras lluvias, la densidad se incrementa enormemente.

Pero la reducción de población propia de la estación seca no es la única variable que tiene implicaciones en este sentido, ya que, además, hay que considerar al menos otras dos. Por un lado, la acción de la selección sobre determinadas inversiones, o sobre combinaciones de inversiones, que afecta a la diferente proporción de sus formas cromosómicas. Por otro lado, también deben tenerse en consideración los eventos de especiación que están teniendo lugar dentro de la especie *A. gambiae sensu stricto*, con el proceso de especiación incipiente de las formas moleculares *M* y *S* y, además, la relación de esta especie con el resto de especies del Complejo, con las que puede mantener cierto grado de flujo génico. Toda esta complejidad, que rodea a la estructura propia de las poblaciones de *A. gambiae*, debe ser atendida a la hora de discutir sobre el tamaño efectivo poblacional de la especie.

Para *A. gambiae* y *A. arabiensis* se realizaron varias estimas independientes usando métodos tanto directos como indirectos (Donnelly, Licht & Lehmann, 2001; Taylor et al., 2001). Estimas indirectas (genéticas) de $N_e$, basadas en la variación temporal de microsatélites o de inversiones cromosómicas, están en el orden de $10^3$ (Taylor et al., 1993; Lehmann et al., 1998; Simard et al., 2000). Todas las estimas de $N_e$ resultaron inconsistentes con la posible existencia de cuellos de botella severos durante la estación seca.

Lehmann et al. (1998) estudiaron la variación temporal (tras siete y nueve años) de las frecuencias de 9 microsatélites de dos poblaciones de *A. gambiae* en Kenia. Las poblaciones estudiadas fueron Asembo, del Oeste de Kenia, y Jego, del Este de Kenia. Estimaron el tamaño efectivo de ambas poblaciones resultando 6359 y 4258, respectivamente para Asembo y Jego. El límite inferior del intervalo de confianza para el 95% fue, respectivamente, 2526 y 1669. Los autores apuntaban que muy posiblemente los

valores reales sean más altos, pues el valor de $N_e$ obtenido estaría muy posiblemente infraestimado al emplearse en el cálculo una tasa de tan solo 12 generaciones por año.

El elevado valor estimado por Lehmann et al (1998) para $N_e$ parece incompatible con la baja densidad de individuos en la estación seca. Los autores proponen que las grandes poblaciones podrían ser mantenidas por individuos ocultos, por ejemplo, adultos que estivan en madrigueras de roedores (Omer & Cloudsley-Thompson, 1970). También proponen que las grandes poblaciones podrían ser mantenidas por una extensiva movilidad de los adultos, resultando en una baja densidad de la *demo* que se encontraría distribuida en una amplia área geográfica. En dicha demo difusa, unos pocos mosquitos sobrevivirían a la estación seca sin reproducirse (estivando), o bien la reproducción podría seguir teniendo lugar en determinados puntos donde el desarrollo de las larvas fuera viable, aunque se necesitan estudios ecológicos para saber cual de estas situaciones ocurre realmente.

El actual $N_e$ para poblaciones del Oeste de África de la especie *A. arabiensis* también se estimó en el orden de $10^3$ (2000 individuos, según Taylor et al., 1993). Esta estima esta basada en el estudio de la variación temporal de las frecuencias de las inversiones cromosómicas. Según Lehmann et al. (1998), una estima tan alta excluye cuellos de botella anuales durante el periodo de estudio. Los resultados obtenidos para *A. gambiae* y *A. arabiensis* revelan que la ecología de estas especies durante la estación seca no es muy conocida.

El elevado valor que fue estimado para el $N_e$ en Kenia es consistente con la mayoría de los estudios que concluyeron que no existía diferenciación genética entre poblaciones separadas por 50 kilómetros (Lehmann et al., 1997; Kamau et al., 1998). Según Lehmann et al. (1998), bajo el modelo de islas de Fisher-Wright, estos resultados sugieren que el área geográfica asociada a una población en el Oeste de Kenia es mayor de 50 kilómetros de diámetro. Esta área sería congruente con la existencia de la "*demo* difusa" descrita.

La distribución de las poblaciones de *A. gambiae* es más o menos continua a lo largo de grandes extensiones, con solo unas pocas regiones en las cuales las poblaciones no se pueden establecer. Así, el modelo de aislamiento por distancia de Wright (1943) describiría muy bien la relación genética entre poblaciones de *A. gambiae* separadas por enormes distancias, en cuanto a que, asumiendo este modelo, una demo con un $N_e$>1000 resultaría en una diferenciación insignificante a lo largo del rango de distribución de la

especie. Esta predicción sería compatible, por una parte, con la ausencia de diferenciación genética entre localidades separadas por hasta 50 kilómetros y, por otra parte, la baja diferenciación registrada en las comparaciones entre localidades separadas por hasta 6000 kilómetros (Lehmann et al., 1996).

Lehmann et al. (1998) concluyeron de su trabajo que durante el periodo de estudio no se han detectado cuellos de botella, sin embargo, esto desecha la posibilidad de que pudieran existir cuellos de botella recientes. Así, Donelly et al. (2001) detectaron trazas de un cuello de botella histórico en Jego. Es más, estudios previos ya habían adelantado la posibilidad de que hubiera ocurrido un cuello de botella en el Este de Kenia. Factores como la baja diversidad genética de sus poblaciones, la presencia de todos los alelos de los microsatélites estudiados en el este de Kenia (con frecuencias mayores de 5%) en el Oeste de Kenia pero la ausencia de varios alelos del Oeste Kenia en el Este, y la evidencia de que la diferenciación en el ADN mitocondrial y en microsatélites fue generada principalmente por deriva (Lehmann et al., 1998; Lehmann et al., 1999; Lehmann et al., 2000).

Lehmann et al. (2003) discutieron sobre las causas de que haya ocurrido un cuello de botella en Jego pero no en Asembo. Apuntan que en África ocurren repetidamente sequías severas, con sus mayores efectos sobre regiones áridas donde la sequía puede durar seis meses o más. Este clima prevalece en muchas partes del Sureste de África. El impacto de las sequías severas quedó muy bien ejemplificado con la sequía de 1770-1780, que secó totalmente el Lago Rukwa (Tanzania) y el Lago Chiuta (Malawi), y redujo los niveles del Lago Malawi en más de 120 metros, produciendo una hambruna que provocó migraciones humanas. Dichos eventos posiblemente redujeron la disponibilidad de lugares adecuados para la cría de *A. gambiae*, y también el número de hospedadores. Las áreas menos afectadas por las sequías son los bosques ecuatoriales y el cinturón de savana húmeda que los rodea. Por lo tanto, Lehmann et al. (2003) creen que es plausible que, tras largos períodos de sequía, las poblaciones presentes en dichas áreas menos afectadas servirían como fuente para reponer las poblaciones de las regiones más afectadas que sufrirían episodios de cuellos de botella y/o extinción cada pocos cientos de años. Este proceso podría ayudar a mantener la homogeneidad genética de *A. gambiae* a lo largo de su rango de distribución, viéndose reducido el "rango efectivo" (aquel con poblaciones permanentes) a aproximadamente la mitad del rango total. La reducción del rango efectivo sería todavía mayor, si tenemos en cuenta que la sequía también reduciría el bosque

ecuatorial y el cinturón de savana húmeda que lo rodea. Pero, si esto es así, ¿cómo es posible que poblaciones del Sahel como, por ejemplo, Barkedji (Senegal), no muestren signos de haber sufrido cuellos de botella y, sin embargo, sí los sufran poblaciones del Sureste de Kenia como Jego?. El Sahel es un estrecho, a menos de 500 kilómetros de la savana húmeda, una distancia relativamente corta que permite una, más o menos, rápida colonización. Además, el Sahel está poblado por la forma molecular *M* de *A. gambiae*, que es más tolerante a las condiciones secas. Por otro lado, la región Sueste de Kenia se encuentra parcialmente aislada por el Valle del Rift, que actúa como barrera, requiriéndose mucho más tiempo para que ocurra una nueva reposición.

Taylor & Manoukis (2003) estudiaron una población compleja de *A. gambiae* de Banambani (Mali). Como ya se ha explicado, parte de la complejidad de las poblaciones del Oeste africano se debe a la existencia de hasta 5 formas cromosómicas y 2 formas moleculares. En este caso, la población estudiada en Banambani comprendía individuos de las formas cromosómicas BAMAKO, SAVANNA y MOPTI. Pero la estructuración de la población estudiada no solamente se debe a presencia de dichas formas cromosómicas ya que, además, al igual que ocurría en las poblaciones de Asembo y Jego anteriormente comentadas, existe una variación temporal del tamaño de población. En este sentido, existe una variación del tamaño de población tanto interanual como estacional, con una reducción del tamaño de población durante la estación seca de hasta un 5-10% del tamaño de población presente en la estación húmeda (Taylor et al., 2001). El tamaño efectivo de población $N_e$ total estimado para Banambani resultó de 4400. Este trabajo, mucho menos minucioso que el comentado anteriormente (Lehmann et al., 1998), resulta interesante en cuanto a que, nuevamente, el valor estimado para $N_e$ se mantiene en el orden de $10^3$ individuos, valor al que parecen ajustarse las poblaciones de *A. gambiae* en África (Krzywinski & Besansky, 2003).

### I.1.4. La cepa *PEST* de *Anopheles gambiae*

La cepa *PEST* fue escogida para la secuenciación del genoma de *A. gambiae* (Holt et al., 2002) debido a que ya existían clones de dos bibliotecas BAC que habían sido secuenciados y localizados físicamente, *in situ*, en los cromosomas. Además, todos los individuos de la colonia presentaban los reordenamientos estándar, sin polimorfismos en ninguna de las inversiones paracéntricas típicas de las poblaciones naturales y de la mayoría de las colonias (Mukabayire & Besansky, 1996). Por último, esta colonia tenía la mutación ligada al cromosoma *X* denominada *pink eye*, que facilitaría la rápida detección de una hipotética contaminación venida de otras colonias.

La mutación *pink eye* (ojo rosa) se originó en 1951, en una colonia denominada *A. gambiae LPE*, mantenida en el *London School of Hygiene and Tropical Medicine* y procedente de mosquitos originarios de Lagos (Nigeria). En 1986 esta mutación se introdujo en una colonia de *A. gambiae* procedente del oeste de Kenia, mediante el cruzamiento de machos de la cepa *LPE* con hembras descendientes de la colonia salvaje de Kenia (forma cromosómica SAVANNA), seleccionando machos de la F2 de este cruzamiento y cruzándolos después con otras hembras descendientes de la misma colonia salvaje de *A. gambiae* procedente de Kenia. De la F2 resultante de este segundo cruzamiento, se seleccionó una cepa en donde se había fijado la mutación *pink eye*. El mismo esquema de cruzamientos se repitió en 1987, consiguiéndose una cepa de *pink eye* con una composición genómica correspondiente mayoritariamente al citotipo SAVANNA del Oeste de Kenia. Esta nueva cepa, denominada *A. gambiae PE* era polimórfica para las inversiones *2La* (32%) y *2Rbc* (19%). La inversión *2Rbc* es característica de la forma cromosómica MOPTI, lo que indica que la cepa original procedente de Nigeria tenía ese mismo citotipo. Finalmente, Mukabaire & Besansky (1996) seleccionaron de esta cepa *PE* un grupo de nueve familias cuya hembra progenitora, y al menos 20 hembras descendientes, tuvieran fijados los cariotipos estándar. La progenie de estas nuevas familias se juntó para conformar la cepa *PEST* de *A. gambiae*.

La cepa *PEST* de *A. gambiae* venía siendo mantenida en USA y en el Instituto Pasteur de París, pero se extinguió. En la actualidad solo se conserva ADN de algunos individuos, aunque tal vez también se conserven algunos individuos enteros congelados.

## I.2. Elementos Genéticos Transponibles.

### I.2.1. Definición de Elemento Transponible.

Los elementos genéticos transponibles (abreviadamente, TEs), también denominados elementos genéticos móviles, son considerados "secuencias de ADN que tienen la capacidad intrínseca de cambiar su localización dentro del genoma que los contiene" [ver, por ejemplo, Li (1997) y Capy et al., (1998)]. Sin embargo, esta definición, a pesar de ser generalmente aceptada, no respeta dos hechos: el primero es que no todos los elementos transponibles tienen la capacidad de cambiar su localización de forma completamente autónoma, sino que algunos necesitan las enzimas codificadas en otros elementos para conseguirlo, y el segundo es que no siempre son los propios elementos los que cambian su localización, sino que la mayoría de estos son copiados, y las nuevas copias son las que se insertan en otros lugares del genoma. Así, teniendo en cuenta estas aclaraciones, los elementos transponibles pueden ser definidos, de una manera general, como "secuencias de ADN que tienen la capacidad de escindirse de su localización original en el genoma que las contiene y reinsertarse en otra, o de ser copiadas e insertar estas copias de sí mismas en distintos lugares del genoma, bien de manera autónoma o bien empleando las enzimas codificadas en otros elementos transponibles".

Los TEs fueron descubiertos por la Doctora McClintock, estudiando el maíz, durante la década de 1940 (McClintock, 1950) y, desde entonces, se ha venido demostrando su presencia en la mayoría de los seres vivos eucariotas estudiados en detalle, con las únicas excepciones conocidas de *Plasmodium falciparum* y, probablemente, otras tantas especies estrechamente relacionadas. Estos elementos constituyen habitualmente una importante parte del genoma de las especies eucariotas que los contienen, siendo la proporción con relación al DNA total variable entre las especies. Así, constituyen aproximadamente la mitad del genoma humano (Venter et al., 2001) o del ratón (Waterston et al., 2002), un 15% del genoma de *Drosophila melanogaster* (Pimpinelli et al., 1995), un 1-3% en hongos (Daboussi & Capy, 2003) y en determinadas plantas llega a representar más del 80% (SanMiguel et al., 1998).

La transposición de los TEs en los genomas implica la mutación del ADN en el que residen, y esta capacidad para crear variabilidad les ha llevado a desempeñar un importante papel en la evolución (revisión por Kidwell & Lisch, 2001).

### I.2.2. Clasificación y estructura de los TEs de eucariotas.

El primer sistema de clasificación fue propuesto por Finnegan (1989). Según este sistema, atendiendo al modo de transposición, los TEs son clasificados en dos grandes grupos: los elementos de Clase I (o retrotransposones), aquellos cuyo ciclo de transposición implica la obtención previa de nuevas copias del elemento mediante la retrotranscripción de un RNA intermediario, y los elementos de Clase II (o DNA transposones), aquellos con transposición directa, mediante la escisión de su lugar original en el genoma y su posterior reinserción en una nueva localización, sin implicar intermediario de RNA alguno. Genéticamente, al mecanismo de transposición de los retrotransposones se le denomina "copiar y pegar", mientras que al mecanismo empleado por los DNA transposones se le denomina "cortar y pegar". Sin embargo, este primer sistema de clasificación se ha visto desafiado por el descubrimiento de determinados TEs, presentes en bacterias y eucariotas, con mecanismo de "copiar y pegar" pero que no emplean RNA intermediario (Duval-Valentin, Marty-Cointin & Chandler, 2004; Lai et al., 2005), y también por el descubrimiento de TEs no autónomos altamente reducidos denominados *MITEs* (*Miniature Inverted Repeat Transposable Elements*).

Recientemente, se han publicado dos revisiones que intentan unificar y actualizar los criterios de clasificación de los TEs (Wicker et al., 2007; Kapitonov & Jurka, 2008). El sistema de clasificación propuesto por Wicker et al. (2007) incluye los niveles de Clase, Subclase, Orden, Superfamilia, Familia. Al igual que la clasificación de Finnegan (1989), el nivel más alto de clasificación (Clase) divide a los TEs en función de la presencia o ausencia de intermediario de RNA en su ciclo de transposición. El nivel de Subclase es utilizado para diferenciar entre elementos con mecanismo de transposición de "cortar y pegar" y de "copiar y pegar". El nivel de Orden marca las diferencias principales del mecanismo empleado en la inserción. Las Superfamilias, dentro del nivel de Orden, comparten estrategias de replicación y, además, no parece existir conservación nucleotídica entre ellas, pero sí a nivel proteico. Las Superfamilias se organizan en Familias, que se

definen en función del grado de conservación de sus secuencias nucleotídicas, conservación que normalmente se restringe a determinados dominios codificadores.

### I.2.2.1. Elementos de clase I o retrotransposones.

Estos elementos también son denominados retrotransposones o retroelementos, debido a que necesitan la actuación de una reverso-transcriptasa para completar su ciclo de transposición. Todos los TEs de Clase I se movilizan mediante la retrotranscripción de un RNA intermediario. No existe el mecanismo de "cortar y pegar", por lo que, según la clasificación propuesta por Wicker et al. (2007), no hay necesidad de hacer una diferenciación entre Subclases. A su vez, los retrotransposones se clasifican en cinco Órdenes distintos atendiendo a características estructurales, enzimáticas y a la filogenética de las reversotranscriptasas. Estos cinco Órdenes son: los retrotransposones con LTRs, los *LINEs*, los *SINEs*, los elementos tipo *DIRS* y los elementos tipo *Penelope*. Los tres primeros grupos son de sobra conocidos de los sistemas de clasificación anteriores (Capy et al., 1997), pero los grupos *DIRS* y *Penelope* han sido descritos recientemente (Goodwin & Poulter, 2004; Evgen'ev & Arkhipova, 2005).

#### I.2.2.1.1. Orden de los retrotransposones con LTRs.

Los retrotransposones con LTRs son el Orden predominante en plantas, mientras que en animales son menos abundantes. Se trata de elementos que normalmente tienen una longitud de entre 4 y 9kb, aunque pueden tener desde unos pocos cientos de bases hasta las 25Kb del elemento *Ogre* (Neumann, Pozarkova & Macas, 2003). Presentan repeticiones largas terminales directas (abreviadamente, LTRs) cuya longitud varía en función del retrotransposón desde a penas unas 100pb hasta 5Kb, flanqueando una región central que generalmente contiene dos grandes ORFs (*Open Reading Frames*, o pautas de lectura abierta), denominados ORF1 y ORF2, y que se corresponden a las regiones *gag* y *pol* de los retrovirus. Sin embargo, determinados elementos de este Orden presentan un único gran *ORF* que incluye ambas regiones *gag* y *pol*, aunque el número, condición, estructura y codificación de los ORFs puede llegar a ser bastante variable (Tubio, Naveira & Costas, 2005).

*Gag* contiene la información codificada para una poliproteína que será procesada en las proteínas que constituyen la cápside, estructura dentro de la cual tiene lugar la retrotranscripción. *Pol* codifica la información para una poliproteína que será procesada dando lugar a las proteínas implicadas en la transposición del elemento: proteasa (PR), reversotranscriptasa (RT), ribonucleasa H (RH) e integrasa (INT). Los retrotransposones se transcriben en un RNA mensajero policistrónico que es traducido en una poliproteína. Las LTRs contienen tres dominios funcionales, denominados U3, R y U5, y contienen en *cis* todos los elementos reguladores requeridos para el inicio de la transcripción y para la poliadenilación del RNA intermediario formado. La transcripción comienza en el extremo 5´ del dominio central R de la LTR5´ y termina en el extremo 3´ de la región R de la LTR3´. El RNA intermediario formado actúa como RNA mensajero, sirviendo como molde para la traducción de una poliproteína que, posteriormente, será procesada por la proteasa. La traducción del ORF2 puede requerir un deslizamiento del ribosoma, en el caso de que exista un cambio en la pauta de lectura entre el ORF1 y el ORF2. Una vez constituida la cápside, podrá tener lugar la retrotranscripción, empaquetándose en ella los otros RNAs obtenidos (en cada cápside se introducen únicamente dos RNAs). El RNA es entonces copiado a un cDNA por la reversotranscriptasa. La retrotranscripción es un proceso complejo que requiere la intervención de un tRNA, cuyo extremo 3´ actúa como cebador y es complementario a una secuencia en el RNA denominada sitio de unión del cebador (PBS, de *Primer Binding Site*). Posteriormente, la RH degrada el RNA de los híbridos DNA-RNA formados y, finalmente, la INT integra la nueva copia en el genoma celular, completándose así un ciclo de transposición. La inserción de las nuevas copias en el ADN genera una duplicación del ADN diana, denominado TSD (*Target Site Duplication*). El TSD varia entre 4-6pb.

Los retrotransposones con LTRs se clasifican en tres Superfamilias, conocidos como "*Ty1/Copia*", "*Ty3/Gypsy*" y "*Pao-Bel*", aunque se tiende a emplear únicamente los nombres *Copia*, *Gypsy* y *Bel*. La diferenciación entre estas tres Superfamilias atiende a la diferente organización estructural de los dominios proteicos codificados en su región *pol*, así como al grado de homología observada tras comparar las secuencias de sus reversotranscriptasas (Xiong & Eickbush, 1990; Xiong, Burke & Eickbush, 1993). En taxonomía vírica, las Superfamilias *Copia* y *Gypsy* también son denominados, respectivamente, *Pseudoviridae* y *Metaviridae* (Boeke. J et al., 1998). Elementos de las

Superfamilias *Copia* y *Gypsy* han sido identificados en plantas, hongos y animales, lo que indica una gran antigüedad, mientras que por el momento los elementos de la Superfamilia *Bel* se han descrito en insectos, nematodos, equinodermos y cordados, y parecen ausentes de los genomas de plantas y de hongos (Copeland et al., 2005).

En los retrotransposones de la Superfamilia *Gypsy*, la región que codifica para la INT se encuentra a continuación del de la RH (al igual que en los retrovirus), mientras que en la Superfamilia *Copia* presentan el dominio de la INT a continuación del correspondiente a la PR y anterior al de la RT. Además, algunos retrotransposones con LTRs presentan un tercer ORF, denominado *env* y similar al gen que codifica para la envuelta externa en los retrovirus. Se ha constatado que este tercer ORF puede ser funcional, como en el caso del retrotransposón *gypsy* en *Drosophila melanogaster*, dotando a los elementos de propiedades infectivas al poderse formar una envuelta vírica (Kim et al., 1994; Song et al., 1994; Syomin et al., 2001). Análisis filogenéticos realizados empleando las secuencias de las reversotranscriptasas han mostrado que los elementos de la Superfamilia *Gypsy* son más próximos filogenéticamente a los retrovirus que a los elementos de la Superfamilia C*opia* (Xiong & Eickbush, 1990).

#### I.2.2.1.2. Orden de los elementos tipo *DIRS*.

Estos elementos carecen de LTRs. Otra característica principal que diferencia a los TEs de este Orden del de los retrotransposones con LTRs es que codifican para una enzima Tirosín-recombinasa (YR), en lugar de una INT, por lo que no forman TSDs. El sistema de clasificación de Wicker et al. (2007) establece tres Superfamilias dentro de este Orden: *DIRS*, *Ngaro* y *VIPER*. La separación de estos tres grupos está basada en el análisis filogenético de la RT y de la RH de varios elementos conocidos de este Orden (Goodwin & Poulter, 2004).

#### I.2.2.1.2.1. Superfamilia *DIRS*.

Uno de los elementos representativos de este grupo es *DIRS1*, identificado en la ameba *Dictyostelium discoideum* (Cappello, Handelsman & Lodish, 1985). Presenta repeticiones terminales invertidas (ITRs) que flanquean una región central que codifica para *gag*, RT, RH y YR. El extremo 3´ de la región central contiene una secuencia conocida como la región complementaria interna (ICR, *Internal Complementary Region*), que es complementaria a las terminaciones de los extremos del elemento y del cual se cree que juega un papel esencial en el ciclo de replicación. Sin embargo, la presencia de ITRs no es una característica común a todos los elementos de esta Superfamilia. Los elementos *PAT* (de Chastonay et al., 1992), del nematodo *Panagrellus redivivus*, y *Kangaroo* (Duncan et al., 2002), del alga *Volvox carteri*, en lugar de ITRs presentan repeticiones directas con diferentes grados de identidad nucleotídica.

#### I.2.2.1.2.2. Superfamilia *Ngaro*.

Los primeros elementos representativos de este grupo fueron identificados en el pez *Danio renio* (Goodwin & Poulter, 2001; Goodwin & Poulter, 2004). Los extremos de estos elementos presentan repeticiones directas con diferentes grados de identidad nucleotídica que flanquean una región central que codifica para GAG, RT, RH, YR. Además, Goodwin & Poulter (2004) identificaron elementos, y/o remanentes de elementos, de la Superfamilia *Ngaro* en otros animales y en hongos.

#### I.2.2.1.2.3. Superfamilia *VIPER*.

Los primeros elementos de este grupo fueron identificados en *Trypanosoma cruzi* y, por el momento, solo se han identificado elementos de este grupo en especies del género *Trypanosoma*. Presentan una estructura similar a los elementos de la Superfamilia *Ngaro*, presentando repeticiones directas con diferentes grados de identidad nucleotídica que flanquean una región central que codifica para GAG, YR, RT, RH (Wicker et al., 2007). El análisis filogenético de la Tirosín-recombinasa reveló que la región YR de los elementos de la Superfamilia *VIPER* presenta mayor homología con las recombinasas de procariotas

que con la región YR de los elementos de las Superfamilias *DIRS* y *Ngaro* (Lorenzi, Robledo & Levin, 2006).

#### I.2.2.1.3. Orden de los elementos tipo *Penelope*.

Los elementos *Penelope* fueron identificados por primera vez en *Drosophila virilis* (Evgen'ev et al., 1997). Se detectaron elementos de este Orden en genomas de más de cincuenta especies, incluyendo eucariotas unicelulares, hongos, animales y plantas (Evgen'ev & Arkhipova, 2005). Estos elementos codifican para una RT que está más estrechamente relacionada con las telomerasas que con las RT de los retrotransposones con LTRs o de los retrotransposones tipo *LINE*. Además, codifican para una endonucleasa homóloga a las endonucleasas codificadas por intrones y a la proteína UvrC, reparadora del ADN de bacterias. Algunos miembros de este Orden pueden retener intrones. Algunos miembros de este Orden pueden presentan falsas LTRs (pseudo-LTRs).

#### I.2.2.1.4. Orden de los *LINE*.

Los elementos pertenecientes al Orden de los *LINE* (*Long Interspersed Nuclear Elements*) han sido identificados en todos los Reinos de los seres vivos eucariotas. Dentro de este Orden se distinguen principalmente cinco Superfamilias: *R2*, *L1*, *RTE*, *I* y *Jockey*.

Suelen tener una longitud de varias kilobases, pero es frecuente encontrarlos truncados en el extremo 5´, presumiblemente como consecuencia de una finalización prematura de la retrotranscripción (Petrov & Hartl, 1998). Generalmente, estos elementos contienen dos ORFs homólogos a los genes *gag* y *pol* de los retrovirus, aunque carecen del dominio de la proteasa del ORF2. La función del ORF *gag*, en aquellos elementos que la presentan, no está clara. Solamente los miembros de la familia *I* presentan RH. En su extremo terminal 3´, los *LINE* pueden presentar una cola de poli-*A*, una repetición en tandem o una secuencia rica en adeninas. Los *LINE* autónomos codifican, al menos, para una RT y para una nucleasa en su ORF *pol*, necesarias para la transposición (Ostertag & Kazazian, 2005). Los elementos de la Superfamilia *RTE*, con únicamente este ORF, se asemejan a lo que, podría ser, un LINE arcaico.

En lo referente a su ciclo de transposición, son transcritos por una RNA polimerasa II, usando promotores internos, en un *RNA* intermediario que se corresponde con toda la secuencia del elemento (McLean, Bucheton & Finnegan, 1993). La reversotranscriptasa reconoce el extremo 3´ del transcrito e inicia simultáneamente la retrotranscripción, a la que seguirá la integración, usando como cebador una mella en el *DNA* genómico (Luan et al., 1993). Este proceso puede generar copias truncadas a las que les falta el extremo 5´ (Eickbush, 1992).

Los *LINE* varían en cuanto a su prevalencia y su diversidad dentro de los genomas. Predominan sobre los retrotransposones con LTRs en muchos genomas animales. Así, por ejemplo, el retrotransposón *L1* llega a constituir el 17% del total del genoma humano (Venter et al., 2001) frente al 8% de los retrotransposones con LTRs (retrovirus endógenos, principalmente); mientras que, en el mosquito *Anopheles gambiae*, a penas unas 100 familias conforman únicamente el 3% del genoma (Biedler & Tu, 2003). Sin embargo, a pesar de su relativa abundancia con respecto a los retrotransposones con LTRs en los genomas animales, en los genomas de plantas suele ocurrir todo lo contrario, siendo la abundancia de los *LINE* normalmente rara con respecto a los LTR.

#### I.2.2.1.5. Orden de los *SINE*.

Los *SINE* (*Short Interspersed Nuclear Elements*) son generalmente elementos menores de 500pb (80-500pb) y, por su pequeño tamaño, no presentan ORFs que puedan codificar alguna de las proteínas implicadas en su transposición. No se trata de elementos derivados de la deleción de otros, sino que se originan de la retrotransposición accidental de tránscritos procesados por una polimerasa III (Kramerov & Vassetzky, 2005). Presentan un promotor para la polimerasa III, lo que les permite expresarse. Además, para su transposición requieren de las RT codificadas por elementos *LINE* (Kajikawa & Okada, 2002; Dewannieux, Esnault & Heidmann, 2003; Kramerov & Vassetzky, 2005). La transposición de los *SINE* produce TSD de 5-15pb. El extremo 5´ del elemento, lugar donde se encuentra el promotor de la polimerasa III, es empleado en la clasificación de los *SINE* en diferentes Superfamilias y, además, indica el origen del elemento: tRNAs, RNA *7SL* ó RNA 5S. El origen del extremo 3´ de los *SINE*, aunque se sabe que puede proceder de *LINEs*, es generalmente desconocido. Este extremo 3´ puede ser una secuencia rica en

*A*, o bien rica en *AT*, portar repeticiones en tandem de 3-5pb o contener una cola de Poli-*A* (Kramerov & Vassetzky, 2005). El *SINE* mejor conocido es la familia *Alu*, que presenta aproximadamente un millón de copias en el genoma humano (Venter et al., 2001).

### I.2.2.2. Elementos de Clase II o transposones de ADN.

Al igual que los elementos de Clase I, los elementos de Clase II se encuentran en los genomas de casi todas las especies de seres vivos eucariotas. Además, de presentarse en los genomas eucariotas, estos elementos también se encuentran en bacterias (siendo conocidos como *IS*, o *insertion sequences*). Los elementos de Clase II se subdividen en dos Subclases, que se diferencian en el número de hebras de ADN que se cortan durante la transposición, pero en ambas Subclases nunca está implicado un RNA intermediario.

#### I.2.2.2.1. Subclase 1.

Contiene a los DNA transposones con un mecanismo de transposición donde el proceso de replicación implica el corte de ambas hebras del ADN. Incluye dos Órdenes: *TIR* y *Crypton*.

##### I.2.2.2.1.1. Orden *TIR*.

Este Orden comprende los elementos clásicos con mecanismo de transposición de "cortar y pegar" y que presentan repeticiones invertidas terminales (TIRs, del inglés de *Terminal Inverted Repeats*) de 10-500pb. A lo largo de su secuencia, entre ambas ITRs, presentan la información codificada para una transposasa que, como su nombre indica, cataliza la transposición. La transposasa se une a/o cerca de las repeticiones invertidas y también al ADN diana. Después, produce una reacción que separa al transposón de su localización original cortando las dos hembras, siguiéndole otra reacción de integración en una nueva localización. Según el sencillo mecanismo de transposición mencionado, parece que no es posible incrementar el número de copias del elemento en el genoma. Sin embargo, diferentes sucesos o estrategias permiten que ocurra tal incremento. Por ejemplo, el elemento *P* en *Drosophila* se escinde en una sola cromátida, y ese lugar es reparado por

la maquinaria celular tomando como molde la cromátida hermana (Daniels & Chovnick, 1993). Es frecuente, en los transposones de DNA, la existencia de elementos "defectuosos" que no codifican una transposasa activa, pero que pueden ser movilizados gracias a la utilización de transposasas codificadas por otros elementos.

Se conocen nueve Superfamilias que se diferencian en la secuencia de las ITRs y en el TSD: *Tc1-Mariner*, *hAT*, *Mutator*, *Merlin*, *Transib*, *P*, *PiggyBac*, *PIF-Harbinger* y *CACTA*.

#### I.2.2.2.1.1.1. Superfamilia *Tc1-Mariner*.

Estos elementos se encuentran ampliamente representados entre los eucariotas. Presentan una estructura típica que comprende dos TIRs que flanquean una región central que codifica para una transposasa (Shao & Tu, 2001). Las familias de este grupo tienen una inserción preferencial por secuencias TA, generando TSD de secuencias TA.

#### I.2.2.2.1.1.2. Superfamilia *hAT*.

Son elementos que presentan un TSD de 8pb, TIRs relativamente cortas de 5-27pb (Kempken & Windhofer, 2001) y un tamaño total que no rebasa las 4Kb. El nombre se deriva de las iniciales de los elementos *hobo* de *Drosophila* (Calvi et al., 1991), *Ac-Ds* del maíz (Courage et al., 1984) y *Tam3* de la planta *Antirrhinum* (Hehl et al., 1991).

#### I.2.2.2.1.1.3. Superfamilia *Mutator*.

Es un grupo de amplio espectro de distribución, encontrándose en todos los Reinos eucariotas, y diverso (Pritham, Feschotte & Wessler, 2005). Aunque las TIRs pueden llegar a tener varios cientos de pares de bases, también pueden ser muy cortas. Su inserción produce TSD de 9-11pb.

#### I.2.2.2.1.1.4. Superfamilia *Merlin*.

La longitud de las TIRs de estos elementos es variable, desde unas docenas de pares de bases a varios cientos. Presentan un TSD de 8-9pb. Solamente se han identificado miembros de esta Superfamilia en animales y eubacterias (Feschotte, 2004).

#### I.2.2.2.1.1.5. Superfamilia *Transib*.

La transposasa de esta Superfamilia presenta homología con la proteína RAG1, implicada en la recombinación de los genes V(D)J del sistema inmune de vertebrados (Kapitonov & Jurka, 2005). Por el momento, solo han sido identificados representantes de *Transib* en *Drosophila* y en mosquitos (Kapitonov & Jurka, 2003).

#### I.2.2.2.1.1.6. Superfamilia *P*.

Esta Superfamilia, inicialmente identificada en genomas de insectos, ha sido encontrada en metazoos (Hammer, Strehl & Hagemann, 2005) y en el alga *Chlamydomonas reinhardtii* (Jurka et al., 2005). Estos elementos generan TSD de 8pb.

#### I.2.2.2.1.1.7. Superfamilia *piggyBac*.

Esta familia se encuentra principalmente en animales (Sarkar et al., 2003), aunque no exclusivamente (Jiang et al., 2005). Tienden a la inserción preferencial en secuencias TTAA.

#### I.2.2.2.1.1.8. Superfamilia *PIF-Harbinger*.

Estos elementos tienen inserción preferencial en secuencias TAA (Jurka & Kapitonov, 2001). Contienen dos ORFs. Uno de los ORFs codifica para una proteína de unión al ADN, el otro codifica para una transposasa.

#### I.2.2.2.1.1.9. Superfamilia *CACTA*.

Uno de sus ORFs codifica para una transposasa, y el otro para una proteína con una función poco clara (Wicker et al., 2003). En plantas, las TIRs terminan en CACTA (a veces CACTG) y generan TSD de 3pb, mientras que en animales y hongos terminan en CCC y se generan TSD de 2pb (DeMarco, Venancio & Verjovski-Almeida, 2006). Las TIRs frecuentemente flanquean series de repeticiones subterminales.

#### I.2.2.2.1.2.Orden *Crypton*.

Este Orden, por el momento muy desconocido, ha sido identificado en hongos. Codifica para una tirosín-recombinasa y no presentan dominio codificador para una RT, lo que sugiere que la transposición de estos elementos ocurre a través de un intermediario de DNA. La transposición supondría la recombinación entre una molécula circular y el ADN diana (Goodwin, Butler & Poulter, 2003; Goodwin & Poulter, 2004), y requeriría el corte de ambas hebras de ADN, motivo por el cual a los elementos de este Orden se les incluye en la Subclase 1. No presentan ITRs, pero parecen generar TSD como resultado de la recombinación e integración.

### I.2.2.2.2. Subclase 2.

Contiene a los DNA transposones con un mecanismo de transposición donde el proceso de replicación no implica el corte de ambas hebras del ADN. La replicación de estos transposones, con transposición de "copiar y pegar", implica el desplazamiento de una de las hebras del ADN solamente. Su clasificación dentro de los elementos de Clase II refleja la ausencia de intermediarios de ARN, pero no necesariamente una relación filogenética próxima con los elementos de Clase I por este hecho.

#### I.2.2.2.2.1. Orden *Helitron*.

Los elementos del Orden *Helitron* parecen replicarse a través de un mecanismo de rodamiento circular, en donde solamente una de las hebras es cortada (Kapitonov & Jurka,

2001). No se producen TSD. Los extremos de estos elementos están definidos por secuencias TC o CTRR (donde R es una purina).

#### I.2.2.2.2.2. Orden *Maverick*.

Los TEs de este Orden también son conocidos como *Positrons*. Estos elementos son muy grandes, comprendiendo 10-20Kb (Feschotte & Pritham, 2005) y están flanqueadas por largas ITRs. Codifican hasta 11 proteínas, pero tanto el número de proteínas codificadas como el orden a lo largo del elemento es variable. Codifican para una DNA polimerasa B y una INT, pero no codifican para una RT. Se han identificado elementos de este Orden en varios eucariotas, pero no en plantas. El mecanismo de transposición supondría la escisión de una de las hebras, seguida de una replicación extracromosómica, tras la que vendría la integración en un nuevo locus (Kapitonov & Jurka, 2005).

### I.2.3. Elementos transponibles y evolución.

#### I.2.3.1. Dinámica evolutiva de los TEs en los genomas eucariotas.

La relevancia de los TEs en la biología de sus hospedadores tenía un interés considerable para Barbara McClintock, descubridora de estos elementos durante la década de 1940. De hecho, ella escogió el nombre de *controlling elements* por la habilidad de estos elementos para "regular la expresión génica de manera precisa" (McClintock, 1956; McClintock, 1984). Esta perspectiva seguía la línea de la suposición predominante en aquel tiempo por la que todos, o la mayoría, de los caracteres de los organismos tenían valor adaptativo, por lo que habrían sido seleccionados favorablemente por la selección natural, es decir, que el mantenimiento de los genes en los genomas a través de las generaciones dependía de la ventaja selectiva que estos proporcionaban al organismo. Más tarde fue creciendo la conciencia de que una gran parte de los genomas estaba constituida por secuencias no codificadoras, mayoritariamente conformadas por ADN de tipo altamente y medianamente repetitivo, y se vio que los TEs constituían una parte importante del ADN medianamente repetitivo de muchos genomas eucariotas. Las características de estas secuencias hacían muy difícil el determinar y asignar alguna función obvia que este ADN repetitivo pudiera proporcionar para el organismo, y sería el desarrollo del neutralismo (Kimura, 1983) lo que proporcionaría la base firme desde la que cuestionar las suposiciones que subyacían a la idea de que los TEs jugaban un papel en gran parte beneficioso en la evolución de sus hospedadores. Pero antes de la publicación de la teoría neutra de la evolución, se publicaron dos trabajos independientes simultáneamente en un mismo número de la revista *Nature*, en 1980, que desafiaban directamente la línea de pensamiento hasta entonces establecida.

En estos artículos (Doolittle & Sapienza, 1980; Orgel & Crick, 1980) se postulaba que la omnipresencia de los elementos repetitivos en los genomas podría ser explicada por mecanismos que no implicaran una ventaja selectiva para el hospedador, concluyendo que el incremento de sus frecuencias y su mantenimiento en las poblaciones naturales puede ser simplemente explicada por la capacidad de estos elementos para replicarse en los genomas que los contienen. Desde entonces, para explicar la persistencia de TEs en los genomas eucariotas se ha venido retomando esta hipótesis, que califica a los TEs como "DNA egoísta" o "DNA parásito". Es más, la conclusión de un trabajo de Hickey (Hickey, 1982) alimentó la teoría del "DNA egoísta", cuando llegó a demostrar teóricamente que es

posible la persistencia de los TEs en las poblaciones naturales aunque supongan una desventaja para los organismos que los hospedan: "*In summary, we can make the following characterization about any DNA sequence that has the ability to self-replicate and transpose within a genome and can thus behave as sexually-transmitted nuclear parasite. Initially, their rate of spread in a population depends on the reproductive success of their hosts but is not equal to it; rather they can spread at twice the rate of the host genes. This is why they may continue to spread provided the reduction in host fitness is less than 50%. However, once they become fixed in a population, the fitness of these transposable elements becomes identical to the fitness of the host; thus their fitness can then be increased only by increasing the fitness of the host. We might consider them to be semi-parasitic genes*".

El modelo propuesto por Hickey (1982) demostraba, en un período en el que arribaba el neutralismo, que el ADN "egoísta" no tenía por qué ser necesariamente neutro. Un hecho significativamente importante en este aspecto es que este ADN puede tener importantes efectos deletéreos sobre el hospedador y, sin embargo, seguir extendiéndose en la población. Años más tarde, Charlesworth Sniegowsky & Stephan (1994) revalorizarían esta visión, apuntando que los TEs se replican a la vez que determinadas mutaciones de los TEs pueden tener importantes repercusiones, reduciendo la *fitness* del organismo. El grado de esta pérdida de fitness causada por los efectos deletéreos de los TEs pudo ser estimado para el elemento *P* en *Drosophila* (Fitzpatrick, 1986; Mackay, 1989; Mackay, Lyman & Jackson, 1992; Currie, Mackay & Partridge, 1998), resultando significativa.

Consistentes con la visión expuesta por Doolittle & Sapienza (1980) y de Orgel y Crick (1980), se propusieron modelos para explicar la dinámica poblacional de los TEs en los genomas eucariotas (revisión realizada por Charlesworth, Sniegowski & Stephan, 1994) concluyendo que los TEs se mantienen en las poblaciones como resultado de un incremento en el número de copias, debido a su actividad transposicional, y a este incremento se opone principalmente la selección, que actuaría de dos formas (Montgomery, Charlesworth & Langley, 1987; Charlesworth & Langley, 1989; Charlesworth, Sniegowski & Stephan, 1994; Charlesworth, Langley & Sniegowski, 1997): 1) selección frente a mutaciones deletéreas debidas a inserciones concretas y 2) selección en contra de reordenamientos cromosómicos, surgidos por recombinación entre elementos

de la misma familia que se encuentran en lugares cromosómicos no homólogos (recombinación ectópica). Además de la transposición y selección, también intervienen otras dos "fuerzas": la escisión y la deriva genética. Sin embargo, las tasas de escisión son mucho menores que las de transposición, sobre todo en los retrotransposones, y no parece tener gran importancia frente a la selección. Por otra parte, la deriva genética solo juega un papel realmente importante en poblaciones con un tamaño efectivo bajo.

El mayor conocimiento que, desde la década de 1980, se ha ido acumulando sobre las interacciones de los TEs en los genomas, ha venido cambiando la visión de los elementos transponibles como estrictamente "parásitos". Es cierto que su movilidad puede resultar deletérea, tal y como se ha mencionado, y que estas secuencias tengan un comportamiento esencialmente parasítico, sin embargo se ha constatado que pueden llegar a suponer un beneficio para las poblaciones y las especies donde se encuentran, y hoy en día es evidente que han venido desempeñando un papel relevante en la evolución de los genomas y las especies eucariotas (Kidwell & Lisch, 2001). A lo largo de la última década se han ido recopilando gran cantidad de evidencias a nivel molecular que han terminado por demostrar, de manera inequívoca, la inmensidad de caminos a través de los cuales los TEs pueden afectar la evolución de las especies que los hospedan. Así, entre otras, los TEs son fuente de nuevos dominios reguladores en *cis* a los genes ya existentes en el genoma hospedador, proporcionando señales de poliadenilación y lugares de *splicing* alternativos, así como nuevos *enhancers*. También intervienen en la adquisición de nuevas funciones celulares, estando implicados en el origen y evolución del sistema inmune de los vertebrados y el mantenimiento de la longitud de los telómeros en *Drosophila*, entre otras funciones celulares conocidas. Además, se ha propuesto a los TEs como causantes del desarrollo de los sistemas de control epigenético en eucariotas.

Al igual que los retrovirus, los TEs dependen de su organismo hospedador para su supervivencia pero, a diferencia de ellos, los elementos móviles no poseen una fase extracelular en su ciclo vital por lo que, en principio, están obligados a una transmisión de tipo vertical. Así, para asegurar la supervivencia de ambos a lo largo del tiempo, es de esperar una coevolución y coadaptación de los TEs con su genoma hospedador, reduciendo el efecto negativo que puede tener la transposición sobre la eficacia biológica de su portador. Teniendo en cuenta que la mayoría de nuevas inserciones tenderán a ser deletéreas para el hospedador, es de interés para ambas partes mitigar esos efectos. Unas de

las estrategias más conocidas de adaptación por parte del hospedador, para mitigar los efectos negativos de la transposición, es el silenciamiento del DNA parásito por mecanismos epigenéticos, como la metilación. Las estrategias de coevolución pueden ser muy variadas, y puede ser que la relación inicial de parasitismo se transforme en una de simbiosis, ya que algunas inserciones podrían llegar a tener un uso esencial y beneficioso para el hospedador (Kidwell & Lisch, 2001).

### I.2.3.2. Impacto estructural y funcional de los TEs en los genomas eucariotas.

El impacto estructural y funcional de los TEs sobre los genomas esta mediado, en general, por dos mecanismos: uno es el proceso de inserción, el otro es la promoción de reordenamientos cromosómicos a través de la recombinación ectópica entre elementos que presentan secuencias homologas. Estos son los dos mecanismos a través de los cuales los TEs inducen variación en los genomas. De estos dos mecanismos, resultan distintos tipos de variación genética que puede ser inducida por los TEs (Kidwell & Lisch, 1997; Kidwell & Lisch, 2001):

#### I.2.3.2.1. Inserción de TEs dentro de exones de genes hospedadores.

Mayoritariamente, este tipo de mutaciones resultan nulas, debido a la baja tolerancia que presentan las secuencias altamente conservadas hacia cualquier tipo de cambio. Sin embargo, aquellas mutaciones que no resultan inviables pueden suponer un incremento de las variantes fenotípicas. Un ejemplo muy recurrido en *Drosophila* es el de la inserción de elementos *P* y *copia* en secuencias exónicas del gen *white*, interrumpiendo la secuencia codificadora de este gen y, por tanto, la producción de pigmento rojo en el ojo (pigmentación propia del alelo salvaje), dando lugar al fenotipo "ojo blanco" que refleja la pérdida de pigmentación (Rubin, Kidwell & Bingham, 1982). Dicha mutación nula puede ser mantenida en cepas de laboratorio, pero su mantenimiento en las poblaciones naturales resultaría muy complicado.

**I.2.3.2.2. Inserción de TEs dentro, o en las proximidades, de las regiones reguladoras de genes hospedadores**.

Probablemente, la inserción de TEs en las regiones corriente arriba al inicio de transcripción de un gen sea el tipo de mutación mediada por TEs con efectos más sorprendentes. Por un lado, la nueva inserción está introduciendo un nuevo inicio de transcripción, y también podría portar su propio promotor, pudiendo producirse un transcrito alternativo que incluyera secuencias propias del TE y del gen. Por otro lado, la nueva inserción podría modificar la expresión del gen alterando sus secuencias de control, bien cambiando la configuración original de las regiones reguladoras del gen, o incluso llegando a aportar nuevas secuencias, que pueden actuar como estimuladores (*enhancers*) que regulen positiva o negativamente la expresión génica.

Actualmente, se conocen gran cantidad de ejemplos de estos cambios reguladores (Medstrand et al., 2005; Thornburg, Gotea & Makalowski, 2006; Volff, 2006; Muotri et al., 2007). Normalmente, solo una pequeña región del TE es la responsable de dichos cambios reguladores, de manera que en aquellos casos favorables la selección natural tenderá a mantener únicamente la pequeña secuencia del TE implicado, lo que muchas veces dificulta su identificación. Un ejemplo de este tipo es el de la inserción del elemento *gypsy* corriente arriba al gen *yellow* en *Drosophila*, que causa una pérdida de expresión de este gen en determinados tejidos (Corces & Geyer, 1991). En *Antirrhinum*, se observó que un elemento *Tam3* se insertó en una región 5´ al gen *niv*, gen que está implicado en la síntesis del pigmento antocianina. La inserción de este elemento implica una regulación negativa de la expresión del gen *niv* (Lister, Jackson & Martin, 1993).

**I.2.3.2.3. Inserción de TEs dentro de intrones de genes hospedadores.**

Las mutaciones de TEs que se insertan en intrones tienen generalmente mayor probabilidad de permanecer a lo largo de un mayor número de generaciones, porque estas mutaciones pasan más desapercibidas a la acción purificadora de la selección natural. Muchas de estas inserciones de TEs son eliminadas durante el procesamiento del ARN mensajero del gen y, por lo tanto, no tienen efectos desfavorables sobre la función propia del gen, auque esto no siempre es así: Nekrutenko & Li (2001), estudiando el impacto de los TEs en genes humanos, detectaron que un 4% (533 genes) de un total de 13799 genes

analizados contenían secuencias propias de TEs dentro de regiones codificadoras. Además, pudieron concluir que aproximadamente el 89.5% de los exones que contienen fragmentos homólogos a TEs correspondían a elementos móviles que originariamente se insertaron en intrones del gen y que posteriormente exonizaron. Esta elevada tasa de exonización, a partir de TEs insertados en intrones, es posible porque muchos TEs portan lugares de *splice* potenciales. Así, por ejemplo, la secuencia consenso para los elementos *Alu* contiene hasta ocho *splice sites* (Makalowski, Mitchell & Labuda, 1994). Además, en algunas ocasiones puede ocurrir que los intrones porten secuencias reguladoras y, aunque el elemento sea eliminado durante el procesamiento del mensajero, su inserción puede afectar a la regulación del gen. Así, por ejemplo, la inserción de elementos *Mu* en el intrón del locus *Knotted* en el maíz induce la expresión ectópica de dicho gen, indicando que dicho intrón porta secuencias normalmente requeridas para reprimir la expresión del gen en determinados tejidos (Greene, Walko & Hake, 1994).

#### I.2.3.2.4. Escisión de TEs del genoma hospedador.

La escisión de TEs es un proceso que también puede ser causante de una variación genética significativa desde un aspecto evolutivo (Wessler, 1988). Esto es debido a que la escisión de un TE de su lugar de inserción original no siempre es preciso. El proceso de escisión puede resultar tanto en la adición de nuevas secuencias como en la pérdida de secuencias que lo flanqueaban (Schiefelbein et al., 1988). Un ejemplo con potenciales implicaciones evolutivas es la del transposón *Ascot-1* en el hongo *Ascobolus immersus* (Colot, Hall & Rosbash, 1988), cuya escisión en un gen que codifica para el color de las esporas deja como huella unas pequeñas secuencias palindrómicas. Esta escisión puede ocurrir en la meiosis, teniendo una importante contribución potencial en la generación de variabilidad en la línea germinal y, por lo tanto, en la evolución de caracteres fenotípicos.

### I.2.3.2.5. Papel reorganizativo de los TEs en los genomas: Transducción y Recombinación ectópica.

La dinámica de los TEs tiene otro papel relevante en los genomas, además de los cambios funcionales hasta ahora descritos, ya que los TEs pueden promover la movilización de grandes y pequeños segmentos de DNA a través de dos mecanismos: la transducción y la recombinación ectópica.

La transducción es consecuencia de una escisión o transcripción no exacta de los TEs en el proceso de transposición, y que implica que en su paso de un *locus A* hacia otro *B* pueden arrastrar consigo secuencias que se encontraban flanqueando al TE en el *locus A*. Moran, DeBerardinis & Kazazian (1999) demostraron este hecho experimentalmente, mediante elementos *LINE-1* modificados para alcanzar altas tasas de retrotransposición en células humanas, los cuales al movilizarse retrotransponen secuencias flanqueantes 3´ a nuevas posiciones genómicas. Es más, se han encontrado *in vivo* retrotransposiciones de tránscritos *LINE-1* conteniendo secuencias flanqueantes que evidencian la transducción de secuencias genómicas (Holmes et al., 1994; McNaughton et al., 1997; Rozmahel et al., 1997).

El hecho de que los TEs se encuentren distribuidos por todo el genoma, así como el que puedan presentar una elevada homología entre ellos, puede llevar a que se produzcan recombinaciones ectópicas entre TEs, o fragmentos de TEs, que se encuentren en distintos *loci* más o menos distantes en un genoma. Las familias multigénicas son susceptibles de recombinación ectópica, en donde tiene lugar un intercambio de secuencias génicas. Este fenómeno puede ser tanto inter como intra-cromosómico, y ocurre entre dos miembros de la Familia localizados en diferentes regiones cromosómicas. Como resultado de dicho proceso de recombinación, se producen reordenamientos cromosómicos como inversiones, deleciones, translocaciones y duplicaciones. Los efectos de esta reorganización pueden llevar potencialmente a la selección natural a oponerse actuando en contra al incremento en el número de copias (Charlesworth, Langley & Sniegowski, 1997). La recombinación ectópica puede promover el reordenamiento de pequeños o grandes fragmentos cromosómicos, implicando mutaciones a pequeña o gran escala que pueden tener un importante impacto en, por ejemplo, la evolución de las especies biológicas (Mathiopoulos et al., 1998; Caceres et al., 1999) o en el origen de determinadas enfermedades (Deininger & Batzer, 1999). Así, no cabe duda de que los cambios reorganizativo-estructurales

promovidos por los TEs pueden tener también importantes implicaciones funcionales. En este sentido es conocido el ejemplo de la mutación dominante *Antp73b* en *Drosophila* (Schneuwly, Kuroiwa & Gehring, 1987). Esta mutación se debe a la recombinación entre dos elementos *Doc*, uno que se encuentra localizado en el primer intrón del gen *antennapedia* y otro en el primer intrón del gen *rfd* (*responsible for dominant phenotype*), ambos con orientación respectiva opuesta. El resultado de la recombinación entre ambos elementos es la generación de una inversión por la cual el promotor del gen *rfd* ejerce un control sobre el gen *antennapedia* que antes no ejercía. El fenotipo resultante es el desarrollo de patas donde debería haber antenas.

### I.2.3.3. Coadaptación TE-genoma hospedador y domesticación de TEs.

Los TEs dependen de sus hospedadores para sobrevivir, de igual manera que ocurre con los virus. Sin embargo, a diferencia con los virus, la mayoría de los TEs no presentan una fase en su ciclo biológico que se desarrolle independiente del hospedador. Por lo tanto, la coadaptación entre los TEs y los genomas que los hospedan debe jugar un importante papel, de cara a la supervivencia a largo plazo de las familias de TEs y de los genomas. Teniendo en cuenta que los TEs tienen, en esencia, un comportamiento parasítico y que, por lo tanto, la mayoría de las nuevas inserciones tendrán un efecto deletéreo para el hospedador, será necesario para ambas partes que se desarrollen vías para mitigar o eliminar dichos efectos deletéreos. A continuación se describen algunos ejemplos de coadaptación entre TEs y genoma hospedador destacados en la bibliografía:

#### I.2.3.3.1. Inserción preferencial en regiones no codificadoras.

Aunque los TEs han venido siendo identificados en prácticamente cualquier localización en los genomas, su distribución claramente no es aleatoria. La observación de que la mayoría de las inserciones de determinadas familias de elementos en *Drosophila*, hongos y plantas se encuentran en regiones no codificadores de genes y fuera de genes, frente a regiones codificadoras, ha llevado a algunos autores a postular que pudiera existir algún tipo de inserción preferencial en este sentido (Kidwell & Lisch, 1997). Frente a esta

observación, se propone la idea de que dicha desviación pudiera ser explicada por la acción temprana de la selección natural, eliminando las inserciones en dichas regiones codificadoras con mayor eficacia que las inserciones de TEs en regiones no codificadoras de los genes. Por ejemplo, en el genoma compacto de *Saccharomyces* se ha anotado que las familias de retroelementos *Ty1*, *Ty2*, *Ty3* y *Ty4* se sitúan corriente arriba a los genes, y la familia *Ty5* en regiones no transcritas del genoma, como telómeros (Kim et al., 1998). Además de los TEs *Het-A* y *TART*, que se insertan en el extremo de los telómeros de *Drosophila* (Levis et al., 1993; Pardue et al., 1996; Biessmann & Mason, 1997), se ha reportado una inserción preferencial de muchos otros TEs en regiones teloméricas: retrotransposones *copia*-like en *Allium cepea* (Pearce et al., 1996), el retroposón *Zep* de *Chlorella* (Higashiyama et al., 1997) y el retroposón *SART1* de *Bombyx mori* (Takahashi, Okazaki & Fujiwara, 1997).

En otros genomas menos compactos que los de *Saccharomyces*, se han encontrado TEs agrupados también en regiones heterocromáticas. Esto ocurre, por ejemplo, en *Drosophila*. En teoría, esta distribución podría ser parcialmente explicada por un incremento de la "supervivencia" en las regiones heterocromáticas, en comparación con la eucromáticas, como consecuencia directa de la diferente densidad de genes en ambos tipos de eucromatina. Sin embargo, las conclusiones de la experimentación con *D. melanogaster* parecen no soportar la hipótesis que supone que el motivo principal para la acumulación de TEs en la heterocromatina sea la acción de la selección en contra de los efectos de los TEs insertados en la eucromatina, tanto por sus efectos directos (propiamente la mutación que supone la inserción) como los indirectos (debido a recombinación ectópica) (Dimitri & Junakovic, 1999). En contraposición a esta conclusión, existen evidencias de la inserción específica en algunas regiones heterocromáticas de *D. melanogaster*. Por ejemplo, la frecuencia de la mutación letal inducida por el elemento *I* en 13 loci localizados en la heterocromatina proximal del cromosoma *2* de *D. melanogaster*, resultó un orden en magnitud mayor de la mutación en genes eucromáticos en el mismo cromosoma (Dimitri, 1997). Este dato proporciona la evidencia de que los elementos *I* se transponen con una mayor frecuencia en las regiones pericentroméricas del cromosoma *2* que están enriquecidas con heterocromatina.

Un ejemplo mucho más acusado, de agrupamiento local de TEs, se conoce en el genoma del maíz. En este genoma, gran parte de las 240Kb entre dos genes está compuesta

por TEs (SanMiguel et al., 1996), incluso se ha reportado en esa región la inserción de unos TEs dentro de otros. Las tasas de recombinación entre TEs del maíz es menor que a nivel de los genes, sugiriendo que la recombinación entre estos TEs (que tiende a ser deletérea para el hospedador) está suprimida (Civardi et al., 1994). Así, parece que las regiones intergénicas del genoma del maíz representan nichos con dominios específicos para ser ocupados por retrotransposones. Sin embargo, no todos los TEs muestran dicha preferencia de inserción. Así, por ejemplo, los *MITEs*, que se encuentran en un número elevado en el maíz, casi siempre se localizan dentro o cerca de genes (Zhang, Arbuckle & Wessler, 2000). Algo similar ocurre con los elementos *Mu* del maíz (Cresse et al., 1995).

La distribución no aleatoria de estos agrupamientos de TEs en los genomas sugiere que los genomas pueden comprender una variedad de "nichos ecológicos", que pueden ser explotados por los diferentes TEs de maneras distintas. La formación de estos nichos, así como su explotación por los TEs, probablemente sea el resultado de un largo periodo de interacción entre el hospedador y su parásito. Los TEs que muestran preferencia por la heterocromatina podrán ser silenciados más fácilmente por la maquinaria celular (Fanti et al., 1998). Sin embargo, los TEs que parecen evitar su inserción dentro de la heterocromatina (como *Mu*), tendrán muchas más posibilidades de mantener su actividad transposicional, aunque se encontrarán con la consecuente acción de la selección natural, más contundente frente al incremento de nuevas copias de dichas familias de elementos. De hecho, el número de inserciones de la familia *Mu* en el genoma del maíz es bajo en comparación con el de los retrotransposones. Por otro lado, los *MITEs*, paradójicamente son abundantes en el genoma del maíz, aunque este hecho parece deberse al pequeño tamaño que presentan, y casi siempre se encuentran en las porciones no codificadoras de los genes, bien como consecuencia de la acción de la selección o bien porque su patrón de inserción les haya llevado a explotar un nuevo "nicho ecológico" dentro del genoma.

#### I.2.3.3.2. Actividad tejido-específica de algunos TEs.

Un buen ejemplo de la adaptación de un TE a su genoma hospedador es la transposición restringida a la línea germinal de los elementos *P* e *I*. La actividad transposicional está restringida en las células somáticas, suponiendo un beneficio para el hospedador, mientras que la transposición en las células germinales incrementa la

posibilidad de que los elementos puedan ser transmitidos verticalmente, asegurando la supervivencia a largo plazo de la familia de TEs. La represión de la transposición del elemento *P* en las células somáticas ocurre a nivel del procesamiento del ARN. La estrategia tejido-específica parece haber sido adoptada por muchos otros elementos; por ejemplo, la expresión de elementos *LINE-1* parece estar favorecida en células de línea germinal de los genomas de humanos (Singer et al., 1993) y del ratón (Trelogan & Martin, 1995); lo mismo ocurre con la expresión del elemento *IAP* en el ratón, que parece estar esencialmente restringida a la línea germinal masculina (Dupressoir & Heidmann, 1996).

#### I.2.3.3.3. Autorregulación de los TEs.

El término "autorregulación" se refiere a la propiedad de algunos TEs que les permite regular su propia tasa de transposición. La autorregulación implica que la tasa de transposición de una familia de elementos es función decreciente del número de elementos de la misma en el genoma hospedador. Esta regulación propia del número de elementos de una familia, ha sido interpretada como una respuesta a los efectos deletéreos de la transposición (Charlesworth & Langley, 1986).

La transposición de elementos *P* en *D. melanogaster* muestra un sistema de autorregulación propio. No cabe duda de que la transposición de elementos *P* implica una pérdida de fitness significativa para *D. melanogaster* (Fitzpatrick, 1986; Mackay, 1989). Para reducir estos efectos se ha desarrollado un sistema de autorregulación, que consiste en la expresión de proteínas represoras de la transposición que son codificadas por elementos de la misma familia (revisión en Capy et al., 1998). En el maíz se conoce un ejemplo similar, que implica la represión de la transposición del transposón *Spm* a través de la proteína codificada por uno de sus genes (Schlappi, Raina & Fedoroff, 1994; Fedoroff, Schlappi & Raina, 1995).

### I.2.3.3.4. Domesticación de TEs en las células eucariotas.

Las secuencias codificadoras internas de los TEs pueden ser reclutadas por el genoma hospedador y desarrollar una función "útil" en dicho genoma. A la emergencia de esta nueva función, que difiere del reclutamiento de secuencias reguladoras comentado en apartados anteriores, se le conoce como domesticación. A continuación se muestran algunos ejemplos conocidos de domesticación de TEs.

#### I.2.3.3.4.1. Telómeros.

Debido a que las DNA polimerasas son insuficientes para replicar los finales mismos de los cromosomas lineales, se hace necesaria la participación de otra maquinaria para asegurar la completa replicación de estos cromosomas. En la mayoría de los eucariotas este problema está resuelto por las telomerasas. Estas enzimas emplean pequeñas moléculas de ARN para, debido a su actividad retrotranscriptasa, añadir residuos que, de otra manera, se perderían durante la replicación. Debido a la importantísima función de esta enzima, uno espera que el uso de la misma se encuentre extremadamente conservado en todos los eucariotas. Sin embargo, *Drosophila melanogaster* se diferencia en este sentido ya que, en lugar de una telomerasa, "emplea" dos familias de retrotransposones para resolver dicho problema; se trata de las familias *Het-A* y *TART* (Levis et al., 1993; Sheen & Levis, 1994). Al término de cada etapa de replicación del ADN, estos elementos se transponen hasta los términos de cada cromosoma replicado. De esta manera, la longitud de los telómeros puede mantenerse relativamente constante, mediante la continua adición de nuevos retrotransposones.

La presencia de determinados TEs en las regiones teloméricas y subteloméricas no es exclusivo de *Drosophila*, y parece un hecho común del que se tiene conocimiento en otros eucariotas, desde insectos (Takahashi, Okazaki & Fujiwara, 1997) hasta algas (Higashiyama et al., 1997). La explotación de estos nichos por parte de los TEs ha venido siendo atribuida al reducido efecto deletéreo que puede suponer para las células la inserción en regiones silenciadas transcripcionalmente. Esta hipótesis sugiere que la habilidad de determinados TEs para explotar las regiones teloméricas haya podido evolucionar (al menos en el caso de *Drosophila*) hacia una interesante relación, con

beneficio mutuo, en la que los TEs han sido co-optados para desempeñar una función celular imprescindible.

#### I.2.3.3.4.2. Reorganización genómica en los protozoos ciliados.

La escisión programada de segmentos de ADN que ocurre durante el ciclo vital de los protozoos ciliados, representa un interesante ejemplo de co-optación de TEs para realizar funciones celulares propias del organismo hospedador.

Cada protozoo ciliado posee un macronúcleo y un micronúcleo. El micronúcleo se encuentra transcripcionalmente inactivo durante la fase vegetativa, mientras que se vuelve transcripcionalmente activo durante la reproducción sexual. El macronúcleo es el elemento activo durante la fase asexual y se encuentra sometido a un proceso de extensivos reordenamientos cromosómicos. Dichos reordenamientos implican la precisa escisión de muchos segmentos de ADN (segmentos conocidos como *internal eliminated sequences*, IES) y su posterior unión en una nueva configuración.

Existe la evidencia de que los TEs conforman una importante fracción de los IES (Klobutcher & Herrick, 1995; Seegmiller et al., 1996; Seegmiller & Herrick, 1998), y se ha propuesto que los TEs estarían directamente implicados en la evolución de este mecanismo de escisión que tiene lugar en los protozoos ciliados (Klobutcher & Herrick, 1997). Es más, se ha sugerido que determinados IES procederían de TEs ancestrales que habrían evolucionado, perdiendo con el tiempo aquellas secuencias innecesarias para el desarrollo de los procesos de escisión que tienen lugar en el macronúcleo.

#### I.2.3.3.4.3. Centrómeros y heterocromatina.

Los centrómeros son las regiones de los cromosomas directamente implicadas en el proceso de segregación de los cromosomas durante la división celular. Generalmente, los centrómeros presentan pocos genes con actividad transcripcional y, además, son heterocromáticos. En muchas especies eucariotas los centrómeros están constituidos en gran parte por TEs y repeticiones simples.

La Familia de elementos *SGM* en *Drosophila guanche* es un interesante ejemplo de co-optación de TEs para formar DNA satélite a gran escala (Miller et al., 2000), ya que la actividad de los elementos de esta Familia en *D. guanche* dio lugar a un DNA satélite que conforma aproximadamente el 10% del genoma, y que se encuentra mayoritariamente en la heterocromatina céntrica.

En mamíferos existen evidencias que sugieren que la co-optación de TEs estaría desempeñando un papel más activo en la función centromérica. En estos vertebrados, las proteínas CENP-B, unas de las encargadas de unir motivos repetitivos dentro de las regiones centroméricas, están claramente relacionadas con el TE *pogo* de *Drosophila* y el transposón *Tigger* de humanos (Earnshaw et al., 1987). De hecho, la repetición terminal invertida del transposón *Tigger* presenta similitud con la secuencia reconocida por las proteínas CENP-B.

**I.2.3.3.4.4. Sistema inmune de vertebrados.**

La diversidad de los anticuerpos y de los receptores T de muchos vertebrados está mediado por un proceso que implica la rotura de segmentos de ADN y una posterior unión de los mismos en una nueva configuración, a través de la recombinación entre segmentos V(D)J. Este proceso ocurre en los linfocitos de cada individuo y es responsable de generar la variabilidad necesaria para responder a un patógeno potencial. Los transposones han sido claves en la evolución de este mecanismo, pues han provisto a los vertebrados de la maquinaria enzimática necesaria para poder desarrollar este nivel de hipervariabilidad somática.

La recombinación V(D)J requiere la actividad de dos enzimas, conocidas como RAG1 y RAG2, que están codificadas, respectivamente, por los genes *rag1* y *rag2*. Investigando la actividad de estas enzimas, resultó llamativo que el mecanismo por el que tiene lugar la recombinación entre los segmentos V(D)J guardaba homologías con el mecanismo de transposición de los transposones (Lewis & Wu, 1997), lo que llevó a proponer que ambos procesos pudieran tener un origen evolutivo común (Spanopoulou et al., 1996). Es más, Agrawal et al. (1998) y Hiom et al. (1998) demostraron que RAG1 y RAG2 pueden llegar a catalizar procesos de transposición *in vitro*, apoyando la propuesta que sugiere que la función original de ambas enzimas era mediar la transposición.

### I.2.3.4. El silenciamiento de los TEs en las células eucariotas. El papel de los TEs en la evolución de determinados mecanismos de represión de la expresión génica.

A lo largo de los últimos años se han venido recopilando evidencias que sugerían que determinados mecanismos de represión de la expresión génica en eucariotas evolucionaron como respuesta adaptativa al comportamiento "parasítico" de los TEs y de los virus (para revisión ver (Matzke et al., 1999; Wolffe & Matzke, 1999; Waterhouse, Wang & Lough, 2001). Como consecuencia del proceso mismo de transposición, los TEs pueden llegar a alterar la estructura y/o la función de los genes del genoma al que parasitan, por lo que las células necesitan un mecanismo para mantener la actividad de estos elementos más o menos controlada. Slotkin & Martienssen (2007) han hecho una exhaustiva actualización sobre la importancia de de los mecanismos transcripcional y post-transcripcional de silenciamiento genético aplicado a los TEs.

#### I.2.3.4.1. Silenciamiento transcripcional de los TEs.

El silenciamiento transcripcional de TEs comprende las modificaciones de la cromatina que pueden suprimir o reducir la actividad transcripcional de los TEs. En este sentido, los mecanismos conocidos implican la metilación de histonas, la metilación del ADN y alteraciones en el empaquetamiento y la condensación de la cromatina.

##### I.2.3.4.1.1. Metilación del ADN.

La metilación del ADN es un mecanismo de silenciamiento de la expresión génica de gran importancia en los vertebrados y en las plantas (Vanyushin, 2006), que consiste en la adición de un grupo metilo en el carbono 5´ de una citosina, denominándose a esta nueva base nucleotídica 5-metil-citosina (abreviadamente, $m^5C$). Actualmente, está muy asumido que este mecanismo de silenciamiento ha evolucionado como una medida de contención frente a la dispersión de los TEs y de virus en los genomas eucariotas (Yoder, Walsh & Bestor, 1997; Waterhouse, Wang & Lough, 2001).

La metilación del ADN como mecanismo de represión de la transcripción ocurre fundamentalmente a nivel de las citosinas en los dinucleótidos CpG (citosina-fosfato-

guanina), aunque también pueden verse metiladas las citosinas en otras disposiciones: así, por ejemplo, en *Drosophila melanogaster*, aunque contiene un nivel muy bajo de citosinas metiladas (Gowher, Leismann & Jeltsch, 2000; Lyko, Ramsahoye & Jaenisch, 2000), esta metilación ocurre principalmente en las citosinas a nivel de los dinucleótidos CpT, aunque también puede darse a nivel de los dinucleótidos CpA y CpC (Lyko, Ramsahoye & Jaenisch, 2000). Así, dado que la metilación de las citosinas puede darse en varios formatos distintos a CpG, se establecieron los términos de "metilación simétrica" y "metilación asimétrica". El primer término hace referencia a que en ambas cadenas complementarias de la hebra de DNA la metilación es simétrica, es decir, la metilación en los dinucleótidos 5´-CpG-3´ es simétrica porque en la cadena complementaria se daría en 3´-GpC-5´. Por el contrario, las citosinas metiladas en CpT y CpA, por ejemplo, son asimétricas.

La metilación simétrica puede darse tanto en plantas como en mamíferos, y puede mantenerse en las nuevas copias de los TEs tras la replicación del DNA, proporcionando un mecanismo para la herencia del silenciamiento. En los ratones, la proteína encargada del mantenimiento de esta metilación del DNA es la metiltransferasa DNMT1. Ha podido observarse que en embriones de ratón deficientes en *Dnmt1* se elevaba la abundancia de tránscritos del retrotransposón *IAP* (Walsh, Chaillet & Bestor, 1998). La metilación asimétrica del DNA también puede darse en los TEs de plantas y, en menor extensión, en mamíferos (Woodcock et al., 1997). Sin embargo, la metilación asimétrica no es heredable, debiendo ser reestablecida a cada TE aparecido *de novo* tras la replicación, sin necesidad de un molde parcialmente metilado para generarse. En el ratón, la metiltransferasa DNMT3 se encarga de la metilación de novo, mayoritariamente a nivel de dinucleótidos CpG (Bourc'his & Bestor, 2004). En *Arabidopsis thaliana*, el homólogo a DNMT1 (DRM2) colabora con otra metiltransferasa (CMT3) para establecer la metilación asimétrica. Sin embargo, aunque la metilación asimétrica del DNA requiera de DRM2 y de CMT3, los TEs serán reprimidos únicamente si está presente MET1, que es la metiltransferasa homóloga a DNMT1 y responsable del mantenimiento de la metilación simétrica (Cao & Jacobsen, 2002). Este hecho indica que la metilación asimétrica del DNA es menos importante que la metilación simétrica de CpG en el mantenimiento del silenciamiento de los TEs.

Recientemente, se ha descubierto que el cambio en el patrón de metilación del DNA en los TEs también puede estar mediado por proteínas implicadas en el empaquetamiento y condensación de la cromatina. Se trata de un grupo de proteínas que, sin ser específicamente metiltransferasas, tienen la capacidad de regular la metilación del DNA, siendo esta la cualidad que les permite actuar sobre el silenciamiento de los TEs. Se trata de proteínas de la superfamilia SNF2 implicadas fundamentalmente en alterar la estructura de la cromatina, para así permitir el acceso de otras proteínas encargadas de la replicación, la transcripción y/o la reparación del DNA. Ha podido constatarse que en las plantas estas proteínas remodeladoras de la cromatina son necesarias para el silenciamiento de TEs. En *Arabidopsis thaliana*, la proteína DDM1 (abreviatura que procede del inglés *Decrease in DNA Methylation*) es requerida para el silenciamiento de TEs y para la condensación de la cromatina (Kato, Takashima & Kakutani, 2004). En el ratón, la proteína Lsh (*lymphoid-specific helicase*), homóloga a la proteína DDM1 de *A. thaliana*, también remodela la cromatina y reprime la actividad de TEs (Yan et al., 2003; Huang et al., 2004). Huang et al. (2004) estudiando los cambios en la transcripción de determinados TEs y de otros genes en mutantes *Lsh*- revelaron que, a pesar de que los tránscritos en los TEs estudiados incrementaban fuertemente su abundancia, no ocurría lo mismo para la mayoría de los genes, cuya transcripción permanecía inalterada. Este hecho sugiere que la actividad remodeladora de Lsh y, por extensión, la de DDM1 podría estar específicamente dirigida para silenciar TEs.

#### I.2.3.4.1.2. Metilación de histonas H3K9.

Las modificaciones en las histonas ocurren en el extremo amino-terminal. Los nucleosomas asociados a los TEs son ricos en histonas H3 metiladas a nivel de su lisina 9 (abreviadamente, H3K9), lo que es un signo de cromatina transcripcionalmente reprimida o inactiva (Gendrel et al., 2002; Martens et al., 2005). Cuando los genes implicados en la metilación de estas histonas se encuentran mutados, existe una reactivación de TEs. Así, por ejemplo, en células embrionarias de ratón se observó que cuando se encuentra mutado el gen *Suv39* (gen de la metiltransferasa de H3K9) se incrementa la abundancia de tránscritos de TEs (Martens et al., 2005).

### I.2.3.4.2. Silenciamiento post-transcripcional de TEs por RNA de interferencia (RNAi).

En el mecanismo de silenciamiento post-transcripcional de TEs, moléculas de RNA de doble cadena (abreviadamente, dsRNA) son procesadas por proteínas de la Familia Dicer en moléculas de RNA más pequeñas, denominadas *small interfering RNA* (abreviadamente, siRNA). Estos siRNA, que tienen una longitud de tan solo 20-30 nucleótidos, son incorporados en un complejo proteico que se encargará de degradar aquellas moléculas de RNA que presenten complementariedad con la secuencia de los correspondientes siRNA. Las proteínas de la Familia Argonauta constituyen el componente catalítico de este complejo proteico encargado de degradar las moléculas de RNA complementario a los siRNA. A este complejo proteico se le conoce como RISC. Se ha demostrado que este mecanismo de silenciamiento de TEs es el más importante en *Caenorhabditis elegans*, por ejemplo para silenciar el transposón de DNA denominado *Tc1* en la línea germinal (Tabara et al., 1999; Sijen & Plasterk, 2003). También se ha visto que mutaciones en proteínas de las Familias Dicer y Argonauta causa la reactivación de TEs en numerosas especies de eucariotas. En la figura 13, se muestran varios posibles escenarios por los que a partir de TEs se pueden generar moléculas de dsRNA que podrían activar el mecanismo de represión post-transcripcional en las células eucariotas.

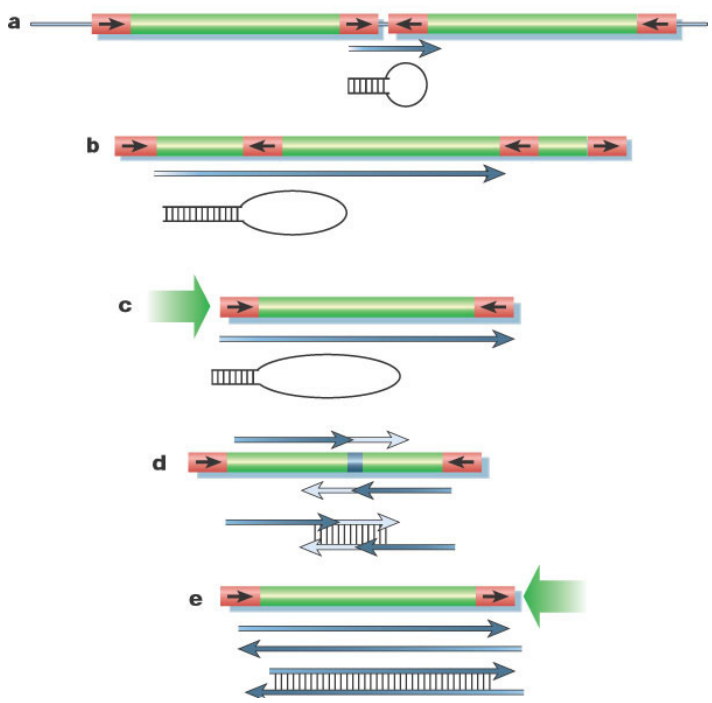



**Figura 13.** Múltiples formas por las cuales los TEs podrían generar secuencias RNAs de doble cadena que podrían activar el mecanismo de represión post-transcripcional en células eucariotas: (a) las LTRs de los TEs de Clase I contienen secuencias promotoras, de manera que dos copias que se inserten próximas y en posición invertida podrían llegar a producir productos de transcripción con secuencias bicatenarias; (b) es frecuente observar un TE embebido en otro TE, en muchos casos en orientación inversa, lo que podría permitir que se generaran RNAs bicatenarios; (c) la inserción de un transposón *TIR* en un punto próximo a un promotor endógeno; (d) la hibridación de ambos productos de expresión de un transposón *TIR* (*Mu*); (e) la inserción de un retrotransposon en un punto adyacente a un promotor endógeno podría dirigir la transcripción en sentido opuesto e inverso a la transcripción normal del retrotransposon, llegándose a producir dos RNAs complementarios que podrían hibridar, produciéndose un dsRNA. Tomada de Waterhouse, Wang & Lough, 2001.

### I.2.3.4.3. Conexión entre los mecanismos de silenciamiento transcripcional de los TEs y el mecanismo post-transcripcional mediado por RNAi.

Aunque el mecanismo de silenciamiento transcripcional de metilación y el post-transcripcional (mediado por RNAi) suelen verse como mecanismos independientes, en donde el primero de ellos supone interacciones DNA-DNA y el segundo RNA-RNA, existen pruebas que indican que ambos procesos, nuclear y citoplasmático, están conectados, de manera que determinadas modificaciones en la cromatina estarían mediadas por moléculas de RNA de interferencia producidas por el mecanismos post-transcripcional

de silenciamiento, que podrían viajar desde el citoplasma hacia el núcleo (Jones, Thomas & Maule, 1998). Sin embargo, el conocimiento científico sobre el RNAi como mediador de modificaciones en la cromatina es, por el momento, incompleto (Slotkin & Martienssen, 2007).

El silenciamiento heterocromático en *Schizosaccharomyces pombe* se debe a un mecanismo de este tipo (Martienssen, Zaratiegui & Goto, 2005) y es empleado como modelo en este sentido: en este hongo, los siRNA generados, a través de proteínas de la Familia Dicer, son incorporados en complejos proteicos denominados RITS, que guiarán la degradación de aquellos tránscritos nacientes que presentan complementariedad con los siRNA previamente generados, mientras todavía están anclados a la RNA polimerasa II durante el proceso de transcripción (Buhler, Verdel & Moazed, 2006). La degradación de estas moléculas nacientes de RNA lleva consigo el marcaje de dicha región de la cromatina para su modificación, a través de la participación de una metiltransferasa de histonas H3K9 y otras proteínas (Irvine et al., 2006). Estas modificaciones consisten en la metilación de histonas H3K9 y, por extensión de este modelo, posiblemente en aquellos organismos en donde el DNA puede ser metilado, las citosinas también puedan ser metiladas por un mecanismo equivalente.

### I.2.4. Los elementos transponibles en *Anopheles gambiae*

Con la secuenciación del genoma de *Anopheles gambiae* (Holt et al., 2002) se ha reportado que los elementos transponibles constituyen alrededor de un 16% de la eucromatina y más de un 60% de la heterocromatina. Los TEs más abundantes parecen ser los elementos de Clase I, sobre todo los *retrotransposones con LTRs*, los *SINEs* y los *MITEs*, aunque también se encuentran representadas la mayoría de las familias de elementos de Clase II.

En el momento de comenzar el proyecto de esta tesis doctoral (año 2002), el número de elementos transponibles de Clase I hasta entonces identificados en el género *Anopheles* era muy bajo: tres elementos representantes del *tipo Ty3/gypsy* [*Aste5* y *Afun1*, (Cook et al., 2000; Hill et al., 2001), *Ozymandias* (Hill et al., 2001)], cinco elementos del

*tipo Ty1/copia* [*Amer3*, *Amer6*, *Amer7* y *Aste7* (Cook et al., 2000; Rohr et al., 2002), *mtanga* (Rohr et al., 2002)], y cinco elementos representantes del *tipo Pao* [*moose* (Biessmann et al., 1999), *Agam10*, *Amer1*, *Aara5* y *Aste12*, (Cook et al., 2000)]. De todos estos TEs, solamente *Ozymandias*, *mtanga* y *moose* fueron identificados en *A. gambiae*. Actualmente, la situación a cambiado significativamente, ya que se han reportado la mayoría de los retrotransposones tipo *LINE* (Biedler & Tu, 2003) y los retrotransposones con LTRs del grupo *Ty3/gypsy* (Tubio, Naveira & Costas, 2005). Además, se han reportado parte de los TEs que representan a los grupos *Pao/Bel* (Marsano & Caizzi, 2005). Los trabajos más detallados hasta ahora publicados han concluido una amplia diversidad de TEs en el genoma de *A. gambiae* (Biedler & Tu, 2003; Tubio, Naveira & Costas, 2005) y una tasa de renovación de retrotransposones significativamente menor que la encontrada en *D. melanogaster* (Tubio, Naveira & Costas, 2005).

La distribución de TEs parece, en principio, consistente con una mayor densidad de éstos en aquellos lugares del genoma con una menor tasa de recombinación. Así, Holt et al. (2002) estimaron que en la eucromatina la presencia de elementos parece mayor cerca de los centrómeros de los cromosomas, mucho menor en las regiones centrales de los brazos, y algo elevada cerca de los telómeros. Además, la densidad de los TEs difiere en función del brazo cromosómico, siendo más alta en el cromosoma X (59 TEs/Mb) y menor en los brazos cromosómicos 3L, 3R, 2L y 2R con 48, 47, 46 y 37 TEs/Mb, respectivamente.

# II. Justificación y Objetivos

## Justificación

*Anopheles gambiae* es el principal vector del protozoo *Plasmodium* que causa la Malaria, una enfermedad que aflige a más de 500 millones de personas y mata a más de 1 millón cada año (Breman, Egan & Keusch, 2001). Esta realidad promovió la secuenciación del genoma del mosquito (Holt et al., 2002), con el objetivo principal de asentar los conocimientos moleculares que permitan desarrollar medidas para intervenir eficazmente sobre la transmisión de la enfermedad.

Los elementos genéticos transponibles (abreviadamente, TEs) son un componente muy importante de muchos seres vivos eucariotas. En el genoma de la especie *Homo sapiens*, por ejemplo, llega a suponer al menos un 45% del tamaño total (Venter et al., 2001). Estos elementos son considerados genéricamente como secuencias de ADN que tienen la capacidad intrínseca de cambiar su localización dentro del genoma que los contiene (Capy, 1998). Su importancia más sobresaliente radica en la capacidad de estas secuencias para generar variabilidad en los genomas como consecuencia de su actividad. Los mecanismos que llevan a los TEs a generar esta variabilidad frecuentemente tienen efectos deletéreos sobre los genomas, lo que les ha llevado a ser considerados secuencias con un comportamiento esencialmente parasítico (Charlesworth, Sniegowski & Stephan, 1994). Sin embargo, esta capacidad para generar variabilidad les lleva a ser muy importantes en la evolución de las especies eucariotas (Kidwell & Lisch, 2001).

De los resultados iniciales de la secuenciación del genoma de *Anopheles gambiae* se reportó que los elementos transponibles conforman aproximadamente un 16% del tamaño total del mismo y más del 60% de la heterocromatina (Holt et al., 2002). Además, se añadía que los retrotransposones con LTRs constituirían el grupo más abundante. La comparación de los patrones del complemento TE entre los distintos genomas pueden proporcionar conclusiones evolutivas muy interesantes, no solo sobre la evolución propia de dichos elementos, sino también sobre la evolución de las especies que los alojan. Por tanto, la secuenciación del genoma de *A. gambiae* representaba un hecho muy importante en el área de la Genética evolutivo-molecular, en cuanto a que permitía obtener patrones comparativos con la otra especie de insectos hasta aquel momento mejor estudiada: *Drosophila melanogaster*.

## Objetivos

El objetivo principal del proyecto de esta tesis doctoral es estudiar la dinámica evolutiva que siguen los retrotransposones del grupo *Ty3/gypsy* en el genoma del mosquito *A. gambiae*, y comparar ésta con el patrón exhibido en otros genomas, principalmente en el genoma de *D. melanogaster* y de *Aedes aegypti*. Para alcanzar este objetivo general, se han marcado los siguientes pasos a seguir:

1. Identificar todas las familias de los retrotransposones del grupo *Ty3/gypsy* en el genoma secuenciado de *A. gambiae* (Holt et al., 2002) y de *Aedes aegypti* (Nene et al., 2007), empleando las herramientas bioinformáticas disponibles para escanear el genoma.

2. Obtener las relaciones filogenéticas de las familias identificadas con los retrotransposones del grupo *Ty3/gypsy* del genoma de *D. melanogaster*, así como con los de otras especies conocidas, como *Aedes aegypti,* agrupándolos en los seis linajes conocidos en insectos según los criterios previamente definidos por Malik & Eickbush (1999).

3. Detectar posibles eventos de transmisión horizontal.

4. Detectar posibles eventos de evolución en mosaico.

5. Caracterizar estructuralmente todas las familias de los linajes del grupo *Ty3/gypsy* del genoma de *A. gambiae*.

6. Identificar todas las inserciones de cada una de las familias del grupo *Ty3/gypsy* en el genoma de *A. gambiae* y clasificarlas según su grado de actividad más probable, atendiendo a parámetros moleculares.

7. Determinar la abundancia relativa de cada uno de los linajes del grupo *Ty3/gypsy* en *Anopheles gambiae* y comparar dicho patrón con el conocido para *D. melanogaster* (Kaminker et al., 2002; Lerat, Rizzon & Biemont, 2003), y con el del mosquito *Aedes aegypti*.

8. Determinar el patrón de renovación de los retrotransposones del grupo *Ty3/gypsy* en el genoma de *A. gambiae*, y compararlo con el patrón conocido para *D. melanogaster*

(Kaminker et al., 2002; Lerat, Rizzon & Biemont, 2003) , y con el del mosquito *Aedes aegypti*.

9. Reportar evidencias moleculares sobre la actuación de las diversas fuerzas evolutivas que pudieran estar implicadas en el patrón exhibido por los retrotransposones del grupo *Ty3/gypsy* en el genoma de *A. gambiae*.

# III. Artículos

# III.1. Artículo 1

## Evolution of the *mdg1* lineage of the *Ty3/gypsy* group of LTR retrotransposons in *Anopheles gambiae.*

*Gene* (2004)

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 330 (2004) 123–131
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

# Evolution of the *mdg1* lineage of the *Ty3/gypsy* group of LTR retrotransposons in *Anopheles gambiae*

Jose Manuel C. Tubío[a], Javier C. Costas[b], Horacio F. Naveira[c,*]

[a] *Departamento de Xenética, Facultade de Bioloxía, Universidade de Santiago de Compostela, Spain*
[b] *Unidade de Mediciña Molecular, INGO, Complexo Hospitalario Universitario de Santiago de Compostela, Spain*
[c] *Departamento de Bioloxía Celular e Molecular, Facultade de Ciencias, Universidade de A Coruña, campus da Zapateira s/n, 15071 A Coruña, Spain*

Received 28 October 2003; received in revised form 22 December 2003; accepted 15 January 2004

Received by D. Finnegan

**Abstract**

So far, only a few retrovirus-like transposable elements (TEs) have been reported in *Anopheles* mosquitoes, although a large fraction of their genomes is made up of these middle repetitive sequences. By screening the *A. gambiae* genome databases, we have found 10 element families belonging to the *mdg1* lineage of the *Ty3/gypsy* group of long terminal repeat (LTR) retrotransposons. These *Anopheles* families constitute a sister clade of the *Drosophila* representatives of this same lineage. According to the phylogenetic reconstruction of their open reading frame (ORF)2 enzymatic domains, the analysis of patterns of nucleotide substitution therein, and the estimation of the age of particular insertions, all these elements must have been active until quite recently, and some of them must be very young. On the other hand, the fact that all these element families are primarily composed of fragmentary copies (mostly solos) or full-length copies with inactivating mutations indicates that their turnover rate has been probably very low. Finally, incongruent phylogenies obtained from different regions of the elements strongly suggest that recombination has played a significant role in their evolutionary history.


*Keywords:* Phylogeny; Recombination; Turnover rate

## 1. Introduction

Transposable elements (TEs) are ubiquitous components of eukaryotic genomes, but differ widely in their abundance. Thus, 4–6% of the euchromatic genome in *Drosophila melanogaster* (Kaminker et al., 2002; Kapitonov and Jurka, 2003), 16% in *Anopheles gambiae* (Holt et al., 2002), and approximately 45% in humans (International Human Genome Sequencing Consortium, 2001) is made up of TE repeats, and the fraction is even larger in many plants (Kumar and Bennetzen, 1999). Both in *Drosophila* and *Anopheles*, LTR retrotransposons constitute the most abundant type of TEs. Insertions of these retrovirus-like elements consist of two LTRs usually encompassing two long open reading frames (ORFs) that correspond to the *gag–pol* regions of retroviruses. Attending to different considerations, two major groups can be distinguished, namely, *Ty1/copia* (*Pseudoviridae*, Boeke et al., 2000a) and *Ty3/gypsy* (*Metaviridae*, Boeke et al., 2000b). Recently, however, a third group with strong bootstrap support and the same structural features as *Ty3/gypsy* has been proposed (*Pao*-like), which includes several retrotransposons related to the *Pao* element from *Bombyx mori* (Cook et al., 2000; Bowen and McDonald, 2001).

The *Ty3/gypsy* group has a widespread distribution in eukaryotes, with representatives from the fungi, animal, and plant kingdoms, notwithstanding their apparently complete absence from birds and the Bdelloidea asexual rotifers, and the presence of a single, highly conserved, orthologous element in mammals (Lynch and Tristem, 2003). Altogether, nine different lineages have been so far identified within the

*Abbreviations:* bp, base pair(s); INT, integrase; kb, kilobase; LTR, long terminal repeat(s); MP, maximum parsimony; Myr, million years; NJ, neighbor joining; ORF, open reading frame; PBS−, minus primer binding site; PR, protease; RNase, ribonuclease; RT, reverse transcriptase; solo, solitary LTR; sORF, short ORF; TEs, transposable elements; TSD, target site duplication.

* Corresponding author. Tel.: +34-981167000; fax: +34-981167065.
*E-mail address:* horaci@udc.es (H.F. Naveira).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.01.012

*Ty3/gypsy* group (Malik and Eickbush, 1999; Bae et al., 2001). Among those that bear envelope domains (endogenous retroviruses), a remarkable diversity of elements has been found, whereas other lineages not bearing *env*-like domains have been relatively disregarded, and thus consist of only one or a few known elements, as it happens with the *mdg1* lineage, presently comprising five representatives, all of them from *Drosophila* (Costas et al., 2001).

The on-going sequencing of the *A. gambiae* genome (Holt et al., 2002) offers an extraordinary opportunity to seek new TEs and gain new insights into their evolution by filling gaps in the phylogeny of their major lineages. In this paper, we report the finding of a subdivision of the *mdg1* lineage of the *Ty3/gypsy* group of LTR retrotransposons in *A. gambiae*, comprising 10 families obtained by screening the genome databases. According to our analyses, these elements must have been active until quite recently, probably with a very low turnover rate, and recombination must have played an important role in their evolution.

## 2. Materials and methods

### 2.1. Screening of TE families of the mdg1 lineage

TBLASTN, from the BLAST server of the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/BLAST) was used to search for sequences homologous to the ORF2 of *pilgrim*, a member of the *D. melanogaster mdg1* lineage (Costas et al., 2001, also known both as *Tabor* and *Wolfman*), in the *A. gambiae* genome (Whole Genome Shotgun project; Holt et al., 2002). A list of the coordinates of all hits showing at least 30% amino acid identity with the query sequence was thus generated. From this list, coordinate intervals were constructed by adding or substracting 6 kilobase (kb) to the highest and lowest coordinate of each hit, respectively. Nucleotide sequences from each interval were then retrieved and examined in greater detail by means of BLAST 2 in the NCBI server, so that the two LTRs of each element could be identified. Possible ORFs were found by sorted three-frame translation of the candidate elements with the aid of BioEdit (available at www.mbio.ncsu.edu/BioEdit/bioedit.html). Different insertions were assigned to the same element family whenever both their ORF1 and ORF2 nucleotide sequences had a pairwise identity of at least 90% over at least 400 base pairs (bp) each. A second run of screening the scaffold database of the *A. gambiae* genome was then carried out, using each of the previously identified elements as probes. A list of coordinate intervals was first generated by parsing the results of BLAST searches using copies resembling as much as possible full-length insertions of each element as query sequence. Intervals were constructed by pooling all hits bearing at least 90% identity over stretches of at least 400 bp that were within 10 kb of each other. A consensus sequence for each TE family was then constructed on the basis of a multiple alignment of different copies retrieved in this way from the genome database.

### 2.2. Number of insertions of each family

To estimate the number of proviral and solo insertions of each family, a second list of coordinate intervals was obtained by BLAST searches with each family consensus, in a similar way as before but excluding short contigs (typically $< 10$ kbp) that could not be adscribed to different insertions with absolute certainty. Intervals bearing identity exclusively to the LTR sequence were assigned to the category of solos, and all the others were assigned to proviral insertions. Whenever possible (i.e., if not truncated either by gaps between contigs or DNA rearrangements), solos were examined for the presence of the characteristic 4 bp target site duplication (TSD) at their ends. Chromosomal locations of the different insertions were obtained from the *A. gambiae* section of the NCBI MapViewer (www.ncbi.nlm.nih.gov/mapview).

### 2.3. Multiple alignments and phylogenetic analyses

After conceptual translation of their consensus nucleotide sequences, the profile alignment option of ClustalX (available at ftp-igbmc.u-strasbg.fr/pub/ClustalX) was used to add the amino acid sequences of elements from *A. gambiae* and a few from *D. melanogaster* to the alignment of the reverse transcriptase (RT), RNase H, and INT regions in the ORF2 from Malik and Eickbush (1999), deposited in the EMBL online database (http://www.es.embnet.org/Services/ftp/databases/embl/align/) under accession no. DS36733, DS36732 and DS36734, respectively. A similar procedure was used to add sequences to the alignment across four conserved domains of the RT protein from Cook et al. (2000). Amino acid sequences corresponding to the ORF1 of the elements obtained after our queries were first aligned using ClustalX (pairwise alignment: gap-opening = 5.0, gap-extension = 10.0; multiple alignment: gap-opening = 5.0, gap-extension = 0.20), and conserved blocks were then selected with the aid of GBlocks (available at monstre1.imim.es/~castresa/Gblocks/Gblocks.html), using relaxed conditions (by setting the minimum number of sequences both for conserved and flank positions to 6, the maximum number of contiguous nonconserved positions to 10, the minimum length of a block to 5, and allowing gaps in less than 50% of the sequences).

Phylogenetic reconstructions of the sequences aligned according to the procedures described above were carried out both by distance (neighbor joining, NJ) and maximum parsimony (MP) methods implemented in the MEGA2.1 package (available at www.megasoftware.net), which give the same weight to all substitutions after removing gaps from the alignments. In MP analyses, we searched for the best trees using the close-neighbor interchange, with default parameter values and random addition of sequences to produce the initial trees. In both MP and NJ analyses,

bootstrapping was used (100 and 1000 replicates, respectively) to assess the support for each internal branch of the trees.

Multiple alignments of nucleotide sequences were obtained from former alignments of amino acid sequences. Synonymous and nonsynonymous substitutions per site were estimated by the method of Nei and Gojobori (1986), implemented in DNaSP (available at www.ub.es/dnasp). Pairwise comparisons of amino acid and nucleotide sequences were done by means of the DotPlot utility of the BioEdit program.

The minus primer binding site (PBS − ) of each element was localized by searching the compilation of tRNA sequences of Sprinzl and Vassilenko (www.uni-bayreuth.de), using sliding-windows of 10 bp at 1 bp steps as probes, starting − 1 bp relative to the 5′ LTR end.

The sequences of TEs first reported in this paper have been deposited in the *A. gambiae* section of Repbase (www.girinst.org/Repbase_Update.html).

## 3. Results and discussion

### 3.1. Phylogenetic relationships

A total of 10 retrovirus-like element families related to the *mdg1* lineage were identified after two rounds of screening the scaffold database of the *A. gambiae* genome. Except for one of them (*GYPSY8_AG*), they are reported here for the first time. Their consensus sequences were reconstructed on the basis of multiple alignments of their different copies scattered throughout the *A. gambiae* genome (*GYPSY8_AG* to *GYPSY17_AG*, Fig. 1; abbreviately, *G8* to *G17* from now on), typically 96–99% identical to the consensus. An alignment across four conserved domains of the RT protein (Xiong and Eickbush, 1990) of these 10 elements, three previously identified *Anopheles* retrotransposons belonging to the *Ty3/gypsy* group (*Afun1*, *Aste11*, and *Ozymandias*; see references in Table 1), at least one representative of each of the nine known lineages of this group, as well as two outgroup *Ty1/copia* elements (Table 1), was phylogenetically analysed both by NJ and MP. The results are shown in Fig. 2. Both kinds of methods placed the 10 elements within the *mdg1* lineage (97–98% bootstrap support) which is split into two reciprocally monophyletic sister clades (*Anopheles* and *Drosophila* representatives), whereas *Ozymandias* is most likely to be a member of the recently discovered *CsRn1* lineage, and *Aste11* is definitely placed into the *gypsy* lineage (in agreement with preliminary phylogenetic analyses carried out by Cook et al., 2000), although it is not an outgroup for the *Drosophila* elements. On the other hand, *Afun1* is not significantly grouped with any of the known lineages.

Several structural features further support the grouping of these novel elements from *A. gambiae* with the representatives of the *mdg1* lineage from *D. melanogaster*. Firstly, all of them use $tRNA^{Arg}$ to prime reverse transcription, contrary to other lineages of the *Ty3/gypsy* group (Terzian et al., 2001), and their 9–12 bp PBS − is located at + 3 relative to the 5′ LTR. Secondly, their INT domain characteristically ends with a GPY/F module, and there is no third downstream ORF encoding an envelope-like domain. Finally, five of them (*G8*, *G9*, *G11*, *G12*, and *G17*) bear a short ORF (sORF) homologous to the sORF2 found in all the *Drosophila* representatives of the *mdg1* lineage.

Some other general properties of these elements are the generation of a 4 bp TSD upon insertion, the absence of a sORF homologous to the sORF1 of *mdg1*, and the presence of two long partially overlapping ORFs, which are out of phase by − 1, a common characteristic of many LTR retrotransposons, that shows the need of a ribosomal frameshifting between the two ORFs. The first one of them should presumably specify a *gag*-like polyprotein, although typical zinc–finger domains could not be found in any of the elements of the *mdg1* lineage (either from *A. gambiae* or *D. melanogaster*). The second gene (*pol*) encodes PR, RT, RNase H, and INT.

### 3.2. Evolutionary history of the mdg1 lineage in A. gambiae

There is increasing evidence that mosaic evolution, caused either by crossing-over or gene conversion, has played a significant role in the patterns of differentiation of retrovirus-like elements (see, for example, Costas and Naveira, 2000). In particular, the *D. melanogaster mdg1* lineage apparently owes much of its diversity to these processes (Costas et al., 2001). The possibility of domain swapping within the *pol* gene was then explored by comparing the phylogenetic reconstructions from conserved regions in either RT, RNase H, or INT (PR was not included in this analysis because its conserved functional domain is very short). Because no cases of disagreement were observed among the three phylogenies (data not shown), it was concluded that interelement recombination had not played a significant role in the evolution of this region, and that the different data sets could be safely combined for more resolutive phylogenetic analyses. Thus, Fig. 3A shows the results obtained by combining the information from RT, RNase H, and INT, which depicts a very clear picture of the evolutionary history of the *mdg1* lineage in *A. gambiae.* There have been three relatively recent element splittings (*G9* vs. *G12*, *G11* vs. *G15*, and *G10* vs. *G17*), which are quite remarkable given the considerable length of the internal branches of the tree. This suggests that a long time interval of evolutionary stasis has been followed by a relatively recent period of accelerated element diversification. This is most patently shown by the branch leading to *G10–G17* which stems from the deepest node of the phylogeny. Assuming a rate of 1.56% synonymous substitutions per synonymous site per million years (Myr; average evolutionary rate for *D. melanogaster* genes, Li,

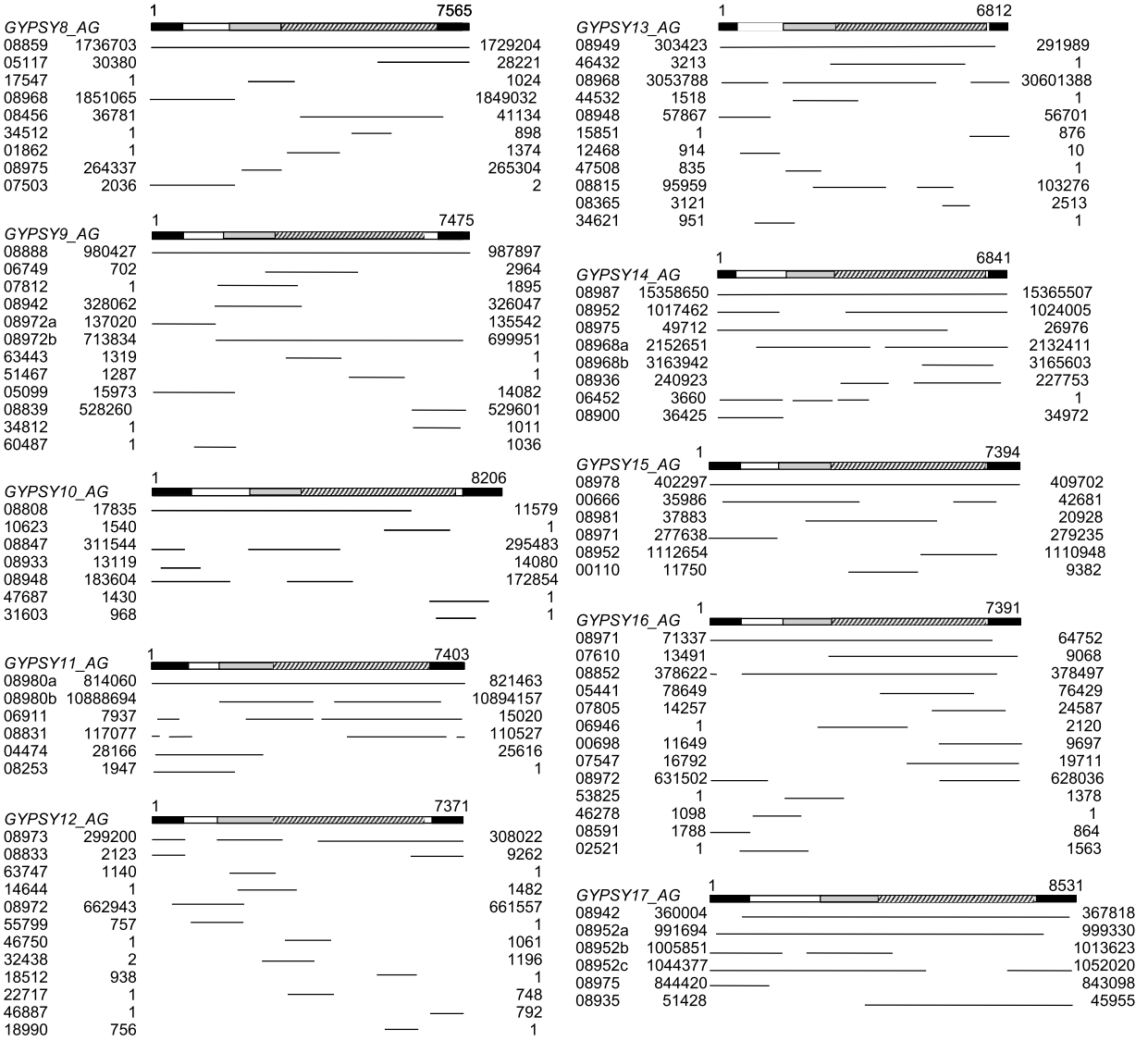



Fig. 1. Reconstruction of the consensus sequences of the 10 elements studied in this paper. Consensus sequences are schematically depicted as rectangles. Black, shaded and striped regions correspond to the LTRs and the two long ORFs, respectively. Total lengths (in bp) are shown above the rectangles. Copies used for reconstruction of the consensus are shown as thick lines beneath the rectangles. Gaps in the lines mark indels. GenBank accession numbers (all entries actually beginning with the string AAAB010) and the corresponding sequence coordinates of TEs are indicated.

1997), and according to our data in Table 2, the three latest splittings can be dated to 14, 12, and 5 Myr ago for *G9–G12*, *G10–G17*, and *G11–G15*, respectively.

In contrast to the ORF2, the analysis of the ORF1 turned out to be very complicated by the low degree of sequence conservation, even at the amino acid level, in some pairwise element comparisons. The 492 positions of the initial alignment were reduced to 220 (44.7%) in 11 conserved blocks after applying the nearly most benign searching conditions of GBlocks. The results of the phylogenetic reconstruction from this final alignment, both by NJ and MP, are shown in Fig. 3B. Three of the two-element clusters found in the previous analysis of ORF2 were also recovered from the analysis of ORF1, namely, *G11–G15*, *G10–G17*, and *G13–G14*, but the topology of both phylogenies as far as *G9*, *G12*, and *G8* are concerned is significantly different. On the one hand, according to the ORF1, *G9* and *G8* constitute a distinct clade (87% bootstrap support by MP), with a considerable distance between them, and remarkably different evolutionary rates along the two external branches. On the other hand, *G12* is not grouped with *G9* but with *G16*, *G14*, and *G13*. Pairwise comparisons of the amino acid sequences of these elements by dot-plots give a clear explanation of these findings (Fig. 4). Thus, the similarity

Table 1
Sequences included in phylogenetic analyses

| Retrotransposon | Host species | Accession no. |
|---|---|---|
| *Ty3/gypsy group* | | |
| *Afun1* | *Anopheles funestus* | – |
| *Aste11* | *Anopheles stephensi* | – |
| *Ozymandias* | *A. gambiae* | AF170018 |
| *mag* | *B. mori* | S08405 |
| *Cer1* | *C. elegans* | U15406 |
| *CsRn1* | *Clonorchis sinensis* | AY013558 |
| *Osvaldo* | *Drosophila buzzatii* | AJ133521 |
| *mdg1* | *D. melanogaster* | X59545 |
| *blood* | *D. melanogaster* | AC005130 |
| *pilgrim* | *D. melanogaster* | AC007146 |
| *Stalker* | *D. melanogaster* | AF420242 |
| *gypsy* | *D. melanogaster* | M12927 |
| *297* | *D. melanogaster* | X03431 |
| *17.6* | *D. melanogaster* | X01472 |
| *412* | *D. melanogaster* | X04132 |
| *mdg3* | *D. melanogaster* | X95908 |
| *Ty3* | *Saccharomyces cerevisiae* | S53577 |
| *Cyclops* | *Vicia faba* | AB007466 |
| *Ty1/copia group* | | |
| *Amer7* | *Anopheles merus* | – |
| *Aste7* | *A. stephensi* | – |

Appropriate references can be found through corresponding GenBank accessions. Sequences with no accession were obtained from Cook et al. (2000).

between *G9* and *G8* is restricted to the first 138 amino acid residues (Fig. 4A), and it is very strong indeed (94% identity), whereas the similarity between *G9* and *G12* extends precisely from this point to the end of ORF1 (Fig. 4B). This seems to be unequivocal evidence of a recent recombination event between *G9* and *G8*, possibly between two RNA genomes packaged within the same virus-like particle (Mikkelsen and Pedersen, 2000). The comparison between *G12* and *G16* (Fig. 4C), showing that their similarity extends throughout the length of ORF1, is consistent with the hypothesis that *G9* got the amino-terminal end of its ORF1 from *G8*.

Other apparent examples of recombination can be found in the LTRs. Multiple alignment of LTR nucleotide sequences of the 10 elements reported in this paper is meaningless due to their extensive lack of homology. As a matter of fact, pairwise comparisons by dot-plots render significant results only in four cases. Two of them reveal prominent incongruencies with the phylogenetic trees formerly built by the analysis of ORF2 and ORF1, namely that, according to the LTRs, *G9* is most similar to *G15* (92.9% nucleotide identity for the first 677 positions of the pairwise alignment as compared to 76.8% for the conserved domains of RT, RNase H, and INT), and *G11* is most similar to *G12* (87.4% identity for the first 580 positions of the alignment in contrast to 76.3% for the coding conserved domains). In both cases, the similarity was interrupted by several indels before reaching the 3′ end of the LTR (Fig. 5A and B, respectively). The other two cases correspond to *G8* vs. either *G9* or *G15*, which produce essentially the same similarity matrix, obviously due to the nearly perfect identity of the LTRs of *G9* and *G15*. In both cases, two homologous tracts were detected, consisting of 125 nucleotides at the 5′ end of the LTR and 174 nucleotides near its 3′ end (data not shown).

### 3.3. Turnover rates and chromosome distribution

The long-term persistence of TEs in the eukaryotic genomes has been explained by models of selfish DNA evolution, which postulate a strong competition among functional copies of a given family within the host genome (Hickey, 1982), although more subtle interactions between different TE families may play a role as well (Leonardo and Nuzhdin, 2002). Natural selection is expected to put limits

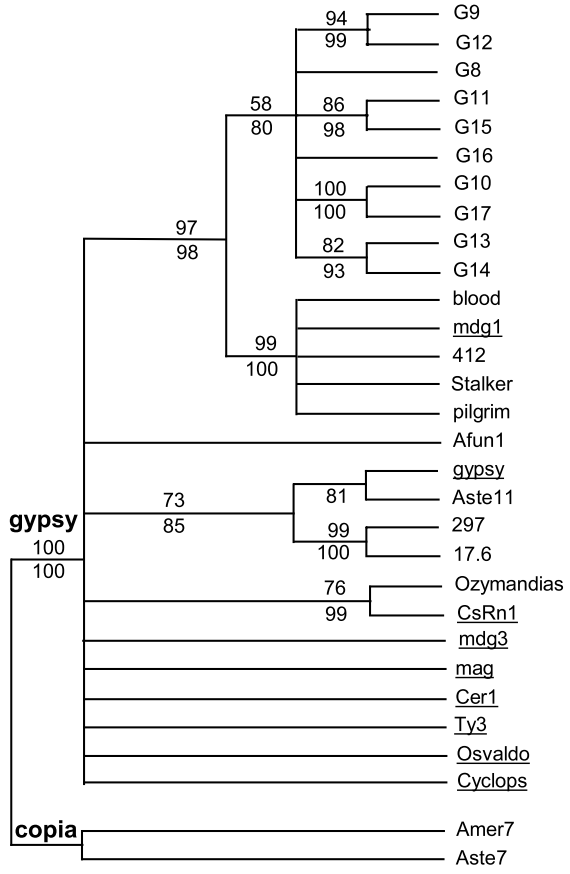



Fig. 2. Bootstrap consensus tree showing the phylogenetic relationships of *Anopheles* retrotransposons within the *Ty3/gypsy* group, based on the amino acid alignment across four conserved domains of the RT coding region. Representative elements of the nine known lineages of this group are underlined. Values indicate bootstrap support of internal branches (MP above, NJ below; only values higher than 50% are indicated). Lengths of all interior branches with 80% or less NJ bootstrap support have been reduced to 0 (condensed tree with 80% cutoff value). Three equally parsimonious trees (876 steps) were obtained. Elements *GYPSY8_A. gambiae* to *GYPSY17_A. gambiae* denoted as *G8* to *G17*.

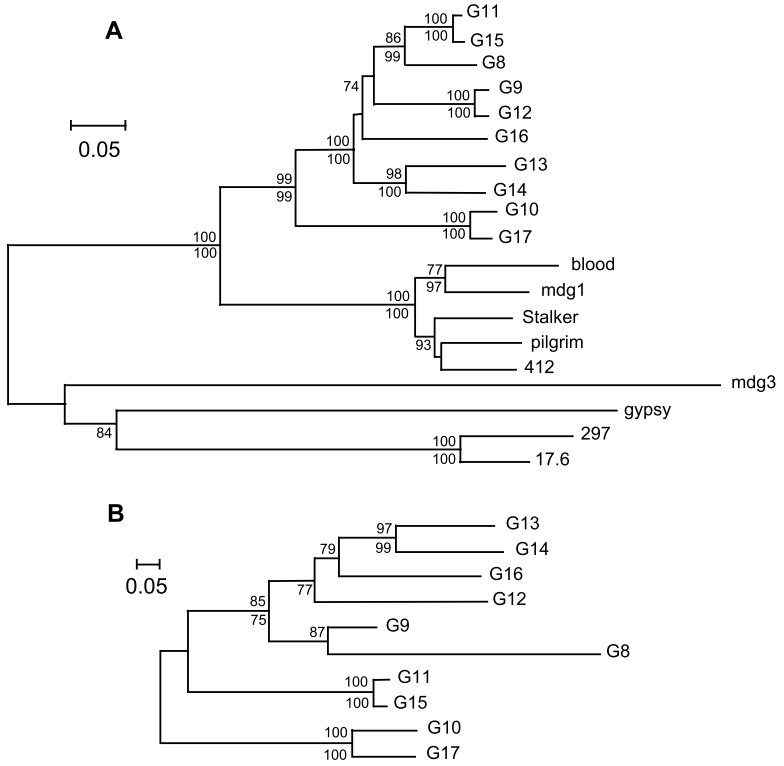



Fig. 3. Fine resolution of phylogenetic relationships of families of the *A. gambiae mdg1* lineage. Both displayed trees were obtained by the NJ method. Values indicate bootstrap support of internal branches (MP above, NJ below; only values higher than 70% are indicated). The scale corresponds to 0.05 amino acid substitutions per site. (A) Based on the alignment of conserved domains of the RT, RNase H and INT coding regions (ORF2). Representative elements of the *D. melanogaster mdg1* and *gypsy* lineages are included for reference. (B) Based on the alignment of 11 conserved amino acid blocks from ORF1 (see Materials and methods for details).

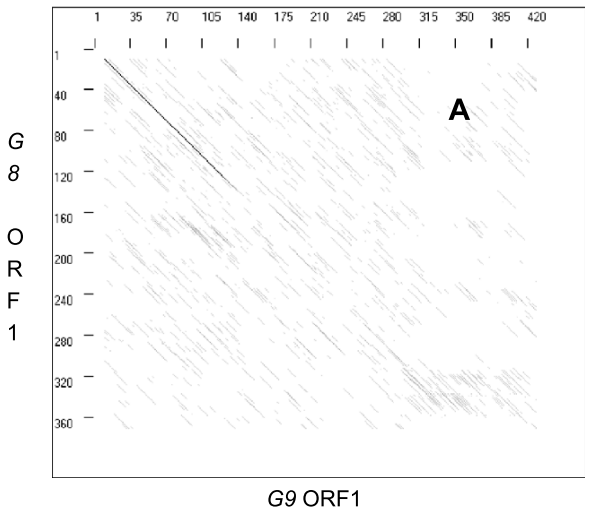
to the total copy number of TEs because of deleterious gene mutations and chromosomal rearrangements promoted by them. Thus, each specific retrotransposon insertion has an indeterminately low probability of fixation, which depends on selective pressure, population size, and recombination frequencies at the surrounding genome (Charlesworth et al., 1997), and it evolves like a pseudogene, acquiring point mutations and indels that eventually render all copies, except the most recent ones, unable to produce functional proteins. Therefore, a retrotransposon family is both the result of a short history of pseudogenic evolution of their existing copies in the host genome and a usually much longer history of purifying selection on their source genes. In agreement with this model, whereas on the one hand, almost all full-length copies so far retrieved from the *A. gambiae* genome (Fig. 1), except for *G11*–08980a, contain frameshift or nonsense mutations in either of their long ORFs, on the other hand, the ratios of synonymous to nonsynonymous substitutions per site for ORF2 domains are generally very high (from 7.0 to 43.1, according to data in Table 2; average = 17.1), indicating that a very strong purifying selection has been acting on these amino acid sites.

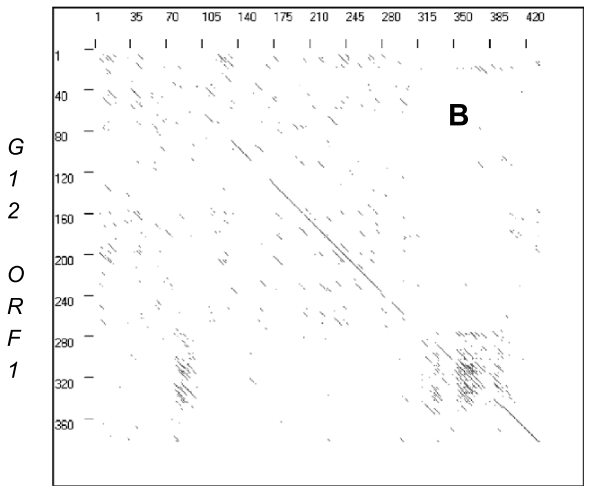
Table 2
Number of pairwise synonymous (above the diagonal) and nonsynonymous (below the diagonal) substitutions per corresponding site for the sequences encoding the ORF2 domains used for Fig. 3A

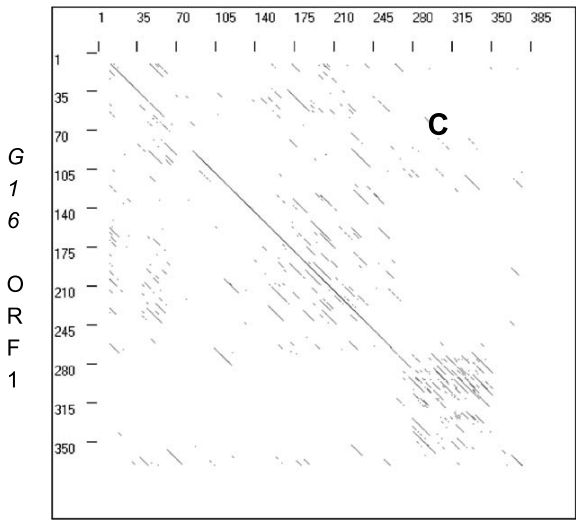
| | *G8* | *G9* | *G10* | *G11* | *G12* | *G13* | *G14* | *G15* | *G16* | *G17* |
|---|---|---|---|---|---|---|---|---|---|---|
| *G8* | | 2.405 | n.a. | 2.791 | 1.980 | n.a. | n.a. | 3.320 | n.a. | n.a. |
| *G9* | 0.136 | | 1.974 | 1.903 | 0.442 | n.a. | n.a. | 1.665 | 3.074 | 1.882 |
| *G10* | 0.229 | 0.248 | | 2.357 | 1.759 | n.a. | n.a. | 2.989 | 2.325 | 0.362 |
| *G11* | 0.075 | 0.113 | 0.236 | | 1.772 | n.a. | n.a. | 0.163 | 2.908 | 3.315 |
| *G12* | 0.137 | 0.015 | 0.250 | 0.119 | | n.a. | n.a. | 1.521 | 3.120 | 2.117 |
| *G13* | 0.175 | 0.101 | 0.266 | 0.141 | 0.168 | | n.a. | n.a. | n.a. | n.a. |
| *G14* | 0.165 | 0.165 | 0.256 | 0.151 | 0.165 | 0.113 | | n.a. | n.a. | n.a. |
| *G15* | 0.077 | 0.117 | 0.242 | 0.008 | 0.122 | 0.147 | 0.155 | | n.a. | 6.464 |
| *G16* | 0.144 | 0.143 | 0.236 | 0.134 | 0.146 | 0.168 | 0.165 | 0.139 | | 2.442 |
| *G17* | 0.221 | 0.243 | 0.022 | 0.234 | 0.243 | 0.255 | 0.248 | 0.237 | 0.227 | |

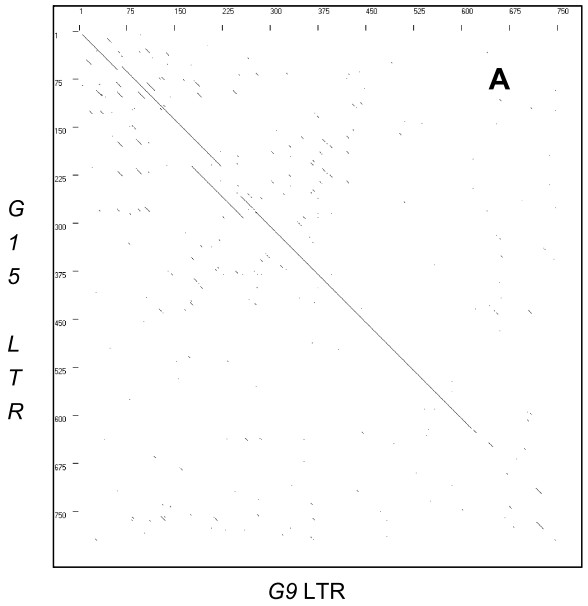
n.a.: not applicable.

The observed number of proviral insertions of the 10 elements of the *mdg1* lineage in the genome of *A. gambiae* is shown in Table 3. The total number of retrieved insertions was 79, and our observations per element range from 4 (*G10* and *G13*) to 16 (*G9*), the average being 7.9. Approximately 60% of these copies can be mapped on the *A. gambiae* chromosomes; the rest of them lie in so far unmapped scaffolds, but this group represents only 16% of the genome assembly (Table S3 in Holt et al., 2002). The available data



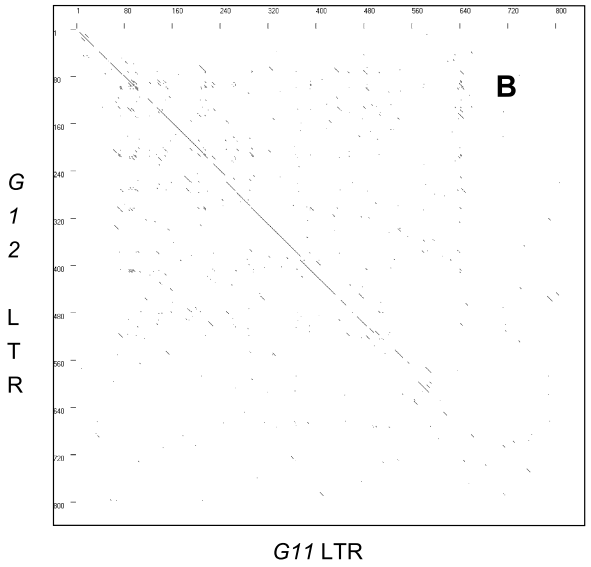




Fig. 5. Pairwise comparisons of nucleotide sequences of the LTRs by dot-plots (window size = 10; mismatch limit = 2). (A) *G9* vs. *G15*. (B) *G11* vs. *G12*. In both cases, the element cited in the first place is on the *x* axis.

show a significant departure ($\chi^2 = 12.4$, $P < 0.005$) from the frequencies expected on the basis of the relative size of each chromosome (9.0%, 47.3%, and 43.7%, for chromosomes X, 2, and 3, respectively; Holt et al., 2002), essentially produced by an excess of insertions on the X chromosome. This finding contrasts with the theoretically expected, and sometimes observed, deficit of TEs on the X chromosome,

Fig. 4. Pairwise comparisons of ORF1 amino acid sequences by dot-plots (window size = 20; similarity matrix = Blosum 62). (A) *G9* vs. *G8*. (B) *G9* vs. *G12*. (C) *G12* vs. *G16*. In all cases, the element cited in the first place is on the *x* axis.

Table 3
Chromosome distribution (either on X, 2 or 3) of the number of copies of the retrovirus-like elements of *A. gambiae* analysed in the present work

| Element | Copy number | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Provirus | | | | | Solitary LTRs | | | | |
| | X | 2 | 3 | NP | Total | X | 2 | 3 | NP | Total |
| *G8* | 3 | 4 | 0 | 3 | 10 | 8 | 8 | 6 | 9 | 31 |
| *G9* | 1 | 4 | 5 | 6 | 16 | 5 | 13 | 27 | 18 | 63[a] |
| *G15* | 3 | 3 | 1 | 8 | 15 | | | | | |
| *G10* | 1 | 1 | 1 | 1 | 4 | 0 | 3 | 7 | 4 | 14 |
| *G11* | 0 | 0 | 3 | 4 | 7 | 0 | 2 | 3 | 2 | 7 |
| *G12* | 0 | 0 | 2 | 3 | 5 | 0 | 0 | 1 | 1 | 2 |
| *G13* | 0 | 2 | 0 | 2 | 4 | 0 | 0 | 1 | 1 | 2 |
| *G14* | 1 | 4 | 1 | 0 | 6 | 1 | 1 | 1 | 1 | 4 |
| *G16* | 1 | 0 | 2 | 3 | 6 | 0 | 2 | 2 | 7 | 11 |
| *G17* | 1 | 3 | 0 | 2 | 6 | 1 | 5 | 9 | 9 | 24 |
| *TOTAL* | 11 | 21 | 15 | 32 | 79 | 15 | 34 | 57 | 52 | 158 |

NP: not placed (unknown chromosomal location).
[a] Solos of *G9* and *G15* are pooled due to their extreme sequence similarity (see Section 3.2 for details).

although recent reports make clear that different TE families may exhibit quite different patterns of chromosome distribution, corresponding to either an excess, a deficit, or no difference at all in the number of X relative to autosome insertions both in *D. melanogaster* (Kaminker et al., 2002; Rizzon et al., 2002; Lerat et al., 2003) and *Caenorhabditis elegans* (Duret et al., 2000).

In addition to proviral insertions, we have also searched for solos. Altogether, 158 insertions of this kind were found, 40% of them identified as either *G9* or *G15* (Table 3), again showing evidence for an increase in the density of X-linked copies ($\chi^2 = 10.8$, $P < 0.005$). Contrary to the overall high proportion of solos (nearly 67% of all insertions) in *A. gambiae*, the sequenced *D. melanogaster* genome harbors very few solos (Kaminker et al., 2002; Lerat et al., 2003). Thus, for the *mdg1* lineage, only three copies have been found (one each for *stalker*, *412*, and *mdg1*). A possible explanation for this striking difference could be that a substantial fraction of what we identified as "solos" in *A. gambiae* actually correspond to retrosequences, produced in similar ways to those giving rise to either SINE-R (after the insertion of a truncated LTR retrotransposon immediately downstream of a strong *Pol III* promoter), or solitary R insertions of HERV-W (after reverse-transcription of the mRNA of an LTR retrotransposon by the RT of a LINE), and not to true solos arising from homologous recombination between the LTRs of the same provirus. To solve this question, we have examined the nucleotide sequences immediately upstream and downstream of each putative solo, looking for the expected 4 bp TSD. Approximately 74% of the insertions included in this category in Table 3 were not truncated in either end of the LTR; 97% of them were actually flanked by 4 bp duplications. Therefore, most, if not all, of these insertions probably correspond to true solos, and the difference between *Anopheles* and *Drosophila* must be for other reasons. Another possibility could be the relative enrichment in fragmented TEs of the sequenced *A. gambiae* genome, given that many scaffolds in the *Anopheles* assembly lie entirely within known heterochromatic regions or extend into centromeres (Holt et al., 2002), whereas nearly all the assembled sequences in *Drosophila* come from the euchromatic part of the genome. However, all the scaffolds in this situation must have been computed as not placed (NP) in Table 3, then representing at most only 33% of all the solos that we have found.

The overall preponderance of fragmentary (including solos) and inactivated full-length copies in all the families reported in this paper indicates that most insertions of these elements have resided for a long time (as pseudogenes) in the *A. gambiae* genome, where they are now probably fixed. At least some of these families (*G8*, *G9*, *G12*, *G14*, and *G15*) are probably still active, however, according to the very high similarity sometimes observed between both LTRs of the same proviral copy (Table 4). In this sense, the observed pattern is similar to that found for most families of LTR retrotransposons in the different plant and animal species analysed to date, except for *D. melanogaster*, where the vast majority of euchromatic element copies have apparently been produced by recent insertions that have had not enough time to diverge (Bowen and McDonald, 2001; Lerat et al., 2003). Differences in turnover rate of TE insertions can be due to a variety of factors, including transposition and excision rates, gene conversion, and, above all, selection against TE insertions, whose effectiveness is mainly affected both by recombination rate (Eickbush and Furano, 2002) and effective population size (Brookfield and Badge, 1997).

The pattern displayed by insertions of the *mdg1* lineage in *A. gambiae* does not necessarily represent a major tendency of turnover rates of TE families in this species.

Table 4
Nucleotide differences between the 5′ and 3′ LTR of the same insertion for the different copies of the *mdg1* lineage so far identified in the *A. gambiae* genome

| Family | GenBank[a] | Position[b] | LTRs | |
|---|---|---|---|---|
| | | | Size[c] | % Divergence |
| *G8* | 08859 | − 1,736,703 | 798 | 0.000 |
| *G9* | 08888 | 980,427 | 751 | 0.000 |
| *G11* | 08980[a] | 814,060 | 811 | 0.123 |
| | 08831 | − 117,077 | 542 | 0.369 |
| *G12* | 08973 | 299,200 | 801 | 0.000 |
| | 08833 | 2123 | 782 | 0.384 |
| *G13* | 08968 | 3,053,788 | 399 | 5,012 |
| *G14* | 08987 | 15,358,650 | 404 | 0.000 |
| | 08952 | 1,017,462 | 404 | 0.000 |
| *G15* | 08978 | 402,297 | 799 | 0.000 |
| *G16* | 08972 | − 631,502 | 698 | 1,433 |
| *G17* | 08952[c] | 1,044,377 | 869 | 1,036 |

[a] GenBank accession numbers of sequences. All entries actually begin with the string AAAB010.
[b] Position of the first nucleotide of the 5′ LTR. A minus sign indicates sequence orientation opposite to the element. See Fig. 1 for further details.
[c] Length in bp of the alignment of the two LTRs excluding gaps.

As a matter of fact, former results obtained by in situ hybridization to polytene chromosomes of the PEST strain of *A. gambiae* (the strain chosen for the *Anopheles* genome project) indicate that there is a significant heterozygosity for insertion sites of other TEs (Mukabayire and Besansky, 1996), which will be probably more akin to the *Drosophila* paradigm (recent, slightly divergent insertions) than the families of the *mdg1* lineage described in this paper (old, very divergent, and probably fixed insertions).

## References

Bae, Y.-A., Moon, S.-Y., Kong, Y., Cho, S.-Y., Rhyu, M.-G., 2001. *CsRn1*, a novel active retrotransposon in a parasitic trematode, *Clonorchis sinensis*, discloses a new phylogenetic clade of *Ty3/gypsy*-like LTR retrotransposons. Mol. Biol. Evol. 18, 1474–1483.

Boeke, J.D., Eickbush, T., Sandmeyer, S.B., Voytas, D.F., 2000a. Family *Pseudoviridae*. In: Regenmortel, M., Fauquet, C., Bishop, D. (Eds.), Virus Taxonomy: Classification and Nomenclature of Viruses, VIIth Report of the ICTV. Academic Press, San Diego, pp. 349–357.

Boeke, J.D., Eickbush, T., Sandmeyer, S.B., Voytas, D.F., 2000b. Family *Metaviridae*. In: Regenmortel, M., Fauquet, C., Bishop, D. (Eds.), Virus Taxonomy: Classification and Nomenclature of Viruses, VIIth Report of the ICTV. Academic Press, San Diego, pp. 359–367.

Bowen, N.J., McDonald, J.F., 2001. *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res. 11, 1527–1540.

Brookfield, J.F.Y., Badge, R.M., 1997. Population genetics models of transposable elements. Genetica 100, 281–294.

Charlesworth, B., Langley, C.H., Sniegowski, P., 1997. Transposable element distributions in *Drosophila*. Genetics 147, 1993–1995.

Cook, J.M., Martin, J., Lewin, A., Sinden, R.E., Tristem, M., 2000. Systematic screening of *Anopheles* mosquito genomes yields evidence for a major clade of *Pao*-like retrotransposons. Insect Mol. Biol. 9, 109–117.

Costas, J., Naveira, H., 2000. Evolutionary history of the human endogenous retrovirus family ERV9. Mol. Biol. Evol. 17, 320–330.

Costas, J., Valadé, E., Naveira, H., 2001. Structural features of the *mdg1* lineage of the *Ty3/gypsy* group of LTR retrotransposons inferred from the phylogenetic analyses of its open reading frames. J. Mol. Evol. 53, 165–171.

Duret, L., Marais, G., Biémont, C., 2000. Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. Genetics 156, 1661–1669.

Eickbush, T.H., Furano, A.V., 2002. Fruit flies and humans respond differently to retrotransposons. Curr. Opin. Genet. Dev. 12, 669–674.

Hickey, D.A., 1982. Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 101, 519–531.

Holt, R.A., et al., 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298, 129–149.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Kaminker, J.S., et al., 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol. 3 (research0084.1–0084.20).

Kapitonov, V.V., Jurka, J., 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc. Natl. Acad. Sci. U. S. A. 100, 6569–6574.

Kumar, A., Bennetzen, J.L., 1999. Plant retrotransposons. Annu. Rev. Genet. 33, 479–532.

Leonardo, T.E., Nuzhdin, S.V., 2002. Intracellular battlegrounds: conflict and cooperation between transposable elements. Genet. Res. 80, 155–161.

Lerat, E., Rizzon, C., Biémont, C., 2003. Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. Genome Res. 13, 1889–1896.

Li, W.H., 1997. Molecular Evolution. Sinauer Associates, Sunderland, p. 191.

Lynch, C., Tristem, M., 2003. A co-opted *gypsy*-type LTR-retrotransposon is conserved in the genomes of humans, sheep, mice, and rats. Curr. Biol. 13, 1518–1523.

Malik, H.S., Eickbush, T.H., 1999. Modular evolution of the integrase domain in the *Ty3/gypsy* class of LTR retrotransposons. J. Virol. 73, 5186–5190.

Mikkelsen, J.G., Pedersen, F.S., 2000. Genetic reassortment and patch repair by recombination in retroviruses. J. Biomed. Sci. 7, 77–99.

Mukabayire, O., Besansky, N.J., 1996. Distribution of *T1*, *Q*, *Pegasus* and *mariner* transposable elements on the polytene chromosomes of PEST, a standard strain of *Anopheles gambiae*. Chromosoma 104, 585–595.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.

Rizzon, C., Marais, G., Gouy, M., Biémont, C., 2002. Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. Genome Res. 12, 400–407.

Terzian, C., Pélisson, A., Bucheton, A., 2001. Evolution and phylogeny of insect endogenous retroviruses. BMC Evol. Biol. I, 3.

Xiong, Y., Eickbush, T.H., 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. 9, 3353–3362.

# III.2. Artículo 2

## Structural and evolutionary analyses of the *Ty3/gypsy group* of LTR retrotransposons in the genome of *Anopheles gambiae.*

*Molecular Biology and Evolution* (2005)

# Structural and Evolutionary Analyses of the *Ty3/gypsy* Group of LTR Retrotransposons in the Genome of *Anopheles gambiae*

*Jose Manuel C. Tubío,* Horacio Naveira,† and Javier Costas‡*

*Departamento de Xenética, Facultade de Bioloxía, Universidade de Santiago de Compostela, Spain; †Departamento de Bioloxía Celular e Molecular, Universidade de A Coruña, Spain; and ‡Unidade de Medicina Molecular, INGO, Complexo Hospitalario Universitario de Santiago de Compostela, Spain

The recent availability of the genome of *Anopheles gambiae* offers an extraordinary opportunity for comparative studies of the diversity of transposable elements (TEs) and their evolutionary dynamics between two related species, taking advantage of the existing information from *Drosophila melanogaster*. To this goal, we screened the genome of *A. gambiae* for elements belonging to the *Ty3/gypsy* group of long-terminal repeat (LTR) retrotransposons. The *A. gambiae* genome displays a rich diversity of LTR retrotransposons, clearly greater than *D. melanogaster*. We have characterized in detail 63 families, belonging to five of the nine main lineages of the *Ty3/gypsy* group. The *Mag* lineage is the most diverse and abundant, with more than 30 families. In sharp contrast with this finding, a single family belonging to this lineage has been found in *D. melanogaster*, here reported for the first time in the literature, most probably consisting of old inactive elements. The *CsRn1* lineage is also abundant in *A. gambiae* but almost absent from *D. melanogaster*. Conversely, the *Osvaldo* lineage has been detected in *Drosophila* but not in Anopheles. Comparison of structural characteristics of different families led to the identification of several lineage-specific features such as the primer-binding site (PBS), the *gag-pol* translational recoding signal (TRS), which is extraordinarily diverse within the *Ty3/gypsy* retrotransposons of *A. gambiae*, or the presence/absence of specific amino acid motifs. Interestingly, some of these characteristics, although in general well conserved within lineages, may have evolved independently in particular branches of the phylogenetic tree. We also show evidence of recent activity for around 75% of the families. Nevertheless, almost all families contain a high proportion of degenerate members and solitary LTRs (solo LTRs), indicative of a lower turnover rate of retrotransposons belonging to the *Ty3/gypsy* group in *A. gambiae* than in *D. melanogaster*. Finally, we have detected significant overrepresentations of insertions on the X chromosome versus autosomes and of putatively active insertions on euchromatin versus heterochromatin.

## Introduction

Eukaryotic transposable elements (TEs) often make up a substantial fraction of the host genome in which they reside. Thus, they constitute 4% to 6% of the euchromatic genome in *Drosophila melanogaster* (Kaminker et al. 2002), 16% in *Anopheles gambiae* (Holt et al. 2002), and 45% in humans (International Human Genome Sequence Consortium 2001).

With the availability of an increasing number of eukaryotic genomic sequences, a primary task in studies of transposon evolution is the characterization of the full transposon complement of sequenced genomes (Holmes 2002; Kaminker et al. 2002). The recently released genome of the Diptera *A. gambiae* (Holt et al. 2002) offers an extraordinary opportunity for comparative studies of TEs diversity and evolutionary dynamics between two related species, taking advantage of the existing information from *D. melanogaster* (Kaminker et al. 2002; Lerat, Rizzon, and Biémont 2003).

The most abundant type of TEs in *Drosophila* is the *Ty3/gypsy* group of long-terminal repeat (LTR) retrotransposons, also referred to as Metaviridae according to virus taxonomy (Boeke et al. 2000). Nine different lineages of this group have been so far identified in different organisms, based on the phylogenetic analysis of their reverse transcriptase (RT), ribonuclease H (RNaseH), and integrase (INT) amino acid domains (Malik and Eickbush 1999; Bae et al. 2001). However, so far, only six of them have been identified in insects, namely *CsRn1, Gypsy, Mag, Mdg1, Mdg3*, and *Osvaldo*. All but *Mag* have been previously detected in *D. melanogaster* (Bae et al. 2001; Kaminker et al. 2002; Kapitonov and Jurka 2003).

Our analysis of the *Mdg1* lineage of *A. gambiae* revealed the existence of 10 different families, mainly consisting of degenerate copies and solitary LTRs (solo LTRs), although some of them also contain very recent, putatively active, insertions (Tubío, Costas, and Naveira 2004). Three additional *Ty3/gypsy* elements have been partially characterized previously; two of them (referred to as *A. gambiae* retrotransposon 1 and *A. gambiae* retrotransposon 2 [Volff et al. 2001]) belong to the *Mag* lineage, whereas the other, *Ozymandias* (Hill et al. 2001), has been assigned to the *CsRn1* lineage (Tubío, Costas, and Naveira 2004). Here, we report our findings on the diversity of the *Ty3/gypsy* group of LTR-retrotransposons in *A. gambiae*. In addition to the recently published study focused on the non-LTR retrotransposons (Biedler and Tu 2003), this work represents an important step towards the characterization of the full set of TEs within the genome of the African malaria mosquito.

## Materials and Methods

### Genome Screening of TE Families of the *Ty3/gypsy* Group

TBlastN (Altschul et al. 1997) was used to search for sequences homologous to the *pol* region of representative elements of each lineage of the *Ty3/gypsy* group in the *A. gambiae* genome (Holt et al. 2002). Specifically, the query sequences included all the well-characterized elements from *D. melanogaster* as well as the *CsRn1*-like

Key words: *Ty3/gypsy*, retrotransposon, *Anopheles gambiae*, *Drosophila melanogaster*.

E-mail: bfcostas@usc.es.

*Mol. Biol. Evol.* 22(1):29–39. 2005
doi:10.1093/molbev/msh251
Advance Access publication September 8, 2004

element within contig AE003787 (positions 212564 to 208162), the retrotransposons *Mag* from the Lepidoptera *Bombyx mori*, *CsRn1* from the Trematoda *Clonorchis sinensis*, *Ty3* from the yeast *Saccharomyces cerevisiae*, *Sushi* from the fish *Fugu rubripes*; *Cyclops* from the plant *Vicia faba*, *Cer1* from the Nematoda *Caenorhabditis elegans*, and *Osvaldo* from *Drosophila buzzatii*. Those hits showing at least 30% amino acid identity over at least 80% of the length of the query sequence were subjected to further analyses, to identify both LTRs of each insertion by means of Blast 2 sequences (Tatusova and Madden 1999). Chromosomal locations of the different insertions were obtained from the *A. gambiae* section of the NCBI MapViewer (www.ncbi.nlm.nih.gov/mapview). Additional BlastN searches were performed using as queries those *A. gambiae* elements identified in this way. This reiterative process (namely, chromosomal allocation of new hits and additional searches) was continued until no new insertions were identified. Different insertions were initially assigned to the same family if they showed at least a stretch of 400 bp of the *pol* region with a pairwise identity of at least 90%. A consensus sequence for each TE family was then constructed by choosing the most frequent nucleotide at each position after manual alignment of the elements with the aid of BioEdit version 5.0.9 (Hall 1999). After construction of family consensus sequences, further refinement of family assignments was carried out by these rules. First, different insertions were assigned to the same family if they presented at least a contiguous stretch of 400 bp of the *pol* region with an identity of at least 90% with the family consensus sequence. Second, insertions were also included in a family if they had a *pol* region shorter than 400 bp but a *gag* region larger than 400 bp with an identity of at least 90% with the family consensus sequence. Finally, those insertions showing (1) *pol* regions shorter than 400 bp with at least 90% identity, (2) *gag* regions shorter than 400 bp with at least 90% identity, or (3) *gag* regions larger than 400 bp with an identity of 85% to 89%, were assigned to the family if they shared, additionally, a minimum of 90% nucleotide identity over at least three quarters the size of the consensus LTR sequence from that family. Solo LTRs were assigned to a specific family if they presented a homology of at least 90% to the consensus LTR sequence.

All the family consensus sequences first reported in this paper have been deposited in the *A. gambiae* section of Repbase Update (http://www.girinst.org/Repbase_Update.html [Jurka 2000]). They were named from *GYPSY18_AG* to *GYPSY72_AG*. The previously discovered element *Ozymandias* (Hill et al. 2001) and the elements *A. gambiae* retrotransposon 1 and *A. gambiae* retrotransposon 2 (Volff et al. 2001) have been renamed as *GYPSY50_AG, GYPSY28_AG*, and *GYPSY55_AG*, respectively, after its full characterization, according to Repbase terminology. The families for which no consensus could be obtained are reported in this paper with the name of the contig or scaffold where a representative sequence was identified.

### Characterization of Insertions

Putative open reading frames (ORFs) were found by sorted three-frame translation of each TE insertion with the aid of BioEdit version 5.0.9. The primer-binding site (PBS) of each element was localized by searching the compilation of tRNA sequences of Sprinzl et al. (1999), using sliding windows of 9 bp at 1-bp steps as probes, starting −1 bp relative to the 5′ LTR end. Individual insertions were considered putatively active if they contained intact ORFs (i.e., without any frameshift or nonsense mutation) and two nontruncated LTRs (i.e., LTRs without indels >10 bp, as compared with the consensus sequence). Those insertions with frameshift mutations, nonsense mutations, or truncated LTRs were classified as inactive insertions. Those insertions with unsequenced gaps but meeting the criteria to be regarded as putatively active based on the analysis of the available sequence were not assigned to a specific activity status. Those insertions bearing identity exclusively to the LTR of a family consensus sequence were considered solo LTRs. Average pairwise divergence between both LTRs from the same element copy and between different copies of the same family were obtained as the proportion of nucleotide differences with the aid of MEGA version 2.1 (Kumar et al. 2002), using the pairwise deletion option.

### Multiple Sequence Alignments and Phylogenetic Analyses

Our phylogenetic analyses were based on the alignment of the seven amino acid domains of the RT defined by Xiong and Eickbush (1990) and the RNaseH and INT domains defined by Malik and Eickbush (1999). The general alignment, available as Supplementary Material online, was obtained in two steps. First, we generated an alignment for each one of the *Ty3*/*gypsy* lineages present in insects using the multiple-alignment mode of ClustalX (Thompson et al. 1997). Each one of the alignments included the consensus sequences of the *A. gambiae* elements of the lineage, the available representative sequences of *D. melanogaster* elements of the lineage, and representative sequences of all the lineages (*Cer1, CsRn1, Cyclops, Gypsy, Mag, Mdg1, Mdg3, Osvaldo*, and *Ty3*). Second, these different alignments were joined together manually, using as guide the representative sequences for each one of the *Ty3*/*gypsy* lineages, common to all the lineage-specific alignments, with the help of BioEdit. For the purpose of phylogenetic analyses, the amino acid motifs of the *D. melanogaster* insertions at genomic sequences AC016130 and AE003787, corresponding to elements belonging to the *Mag* and *CsRn1* lineages, respectively, have been reconstructed by the introduction of gaps to compensate for frameshift mutations.

Phylogenetic relationships between different retrotransposons based on this general alignment were obtained both by distance (neighbor-joining [NJ]) and maximum-parsimony (MP) methods, as implemented in MEGA version 2.1, using the pairwise deletion option. The amino acid distances were computed using the Poisson correction for multiple substitutions and assuming equality of substitution rates among sites. In MP analyses, we searched for the best tree using the close-neighbor interchange, with default parameter values and random addition of sequences to produce the initial trees. In both MP and NJ analyses,

bootstrapping was performed (1,000 replicates) to assess the support for each internal branch of the tree.

### Statistical Analysis of the Distribution of Insertions in the X Chromosome Versus the Autosomes

The equiproportional hypothesis of Montgomery, Charlesworth, and Langley (1987) postulates that the turnover of insertions should occur at equal rates on the X chromosome and the autosomes. Under this hypothesis, the expected ratio of haploid mean copy number of any given family in the X chromosome and the autosomes ($H_X/H_A$) at equilibrium can be obtained by solving the quadratic in $X = H_X/H_A$, after assigning numerical values to the constants in equation (2) of Montgomery, Charlesworth, and Langley (1987), corrected after Langley et al. (1988). We followed the statements of Krzywinski et al. (2004) and assumed that the Y chromosome is entirely heterochromatic, that it constitutes 10% of the haploid genome of a male, and that 975 out of the 8,845 total unmapped scaffolds of the *A. gambiae* genome are most likely to be linked to this chromosome. We also followed Holt et al. (2002) for assumptions on the relative size of each chromosome. All unmapped scaffolds, amounting to roughly 44 Mb, were pooled into a separate category, conceptually equivalent to part of the "heterochromatin" in the models of Montgomery, Charlesworth, and Langley (1987). Finally, after assuming that transposition rates per copy per generation do not differ either between sexes or between heterochromatic and euchromatic insertions, a value of 8.0% was obtained for the expected proportion of elements on the X chromosome under this equiproportional hypothesis. Observed and expected frequencies were compared by means of $\chi^2$ tests.

## Results

### *Ty3/gypsy* Families in the *A. gambiae* Genome

We have identified three known (*Ozymandias*, *A. gambiae* retrotransposon 1, and *A. gambiae* retrotransposon 2) and 127 putative novel families of retrotransposons belonging to the *Ty3/gypsy* group in the sequenced genome of *A. gambiae*, in addition to the 10 families from the *Mdg1* lineage previously reported by our group (Tubío, Costas, and Naveira 2004). Sixty-three of these families, representing those cases where either it was possible to obtain a consensus sequence or there was at least one complete insertion sequence potentially capable of transposition, are characterized below. Those insertions not classified in any of these families are succinctly described in table 1 of Supplementary Material online.

Figure 1 shows the phylogenetic relationships of all the well-characterized families from *D. melanogaster* and *A. gambiae* belonging to the *Ty3/gypsy* group of LTR retrotransposons, based on the alignment of the conserved amino acid domains of RT, RNaseH, and INT; in addition to *D. melanogaster* elements belonging to the *Mag* (AC016130) and *CsRn1* (AE003787) lineages, as well as representatives from other species of each one of the six lineages of the *Ty3/gypsy* group without well-characterized sequences in *D. melanogaster* genome.

The high bootstrap values (>95% in all cases) support an unambiguous assignment of all mosquito elements to each one of five different lineages, namely *CsRn1, Gypsy, Mag, Mdg1*, and *Mdg3*. The only exception is an *A. gambiae* family from the *Mdg1* lineage (*GYPSY54_AG*, undetected in our previous work [Tubío, Costas, and Naveira 2004]) representing a very basal branch in the phylogenetic tree of this lineage. Although its clustering with members of the *Mdg1* lineage is well supported by the NJ method (98% bootstrap value), this is not the case using MP (63% bootstrap value). In addition to the well-supported phylogenetic relationships clearly defining each lineage, all the elements of each one of the lineages have distinctive structural characteristics, presented in figure 2. There is also some structural variation within lineages. Thus, families from the *CsRn1* lineage show two alternative translational recoding signals (TRSs), some families from the *Gypsy* lineage differ in the presence or absence of an additional ORF encoding the *env* protein, and a few families from the *Mag* lineage have a particular TRS (fig. 2).

### The *CsRn1* Lineage

This lineage was first described in trematodes and has also been detected in *D. melanogaster* and *A. gambiae* (Bae et al. 2001; Hill et al. 2001). We have identified 21 putative families, although we have only been able to obtain a consensus sequence for seven of them because of the low number of insertions available for the other families.

In contrast to *A. gambiae*, the *CsRn1* lineage appears to be poorly represented in the *D. melanogaster* genome. Our Blast searches to evaluate this observation revealed the existence of only one family (referred to as AE003787DM in the phylogeny) with five insertions in the fly genome: three solo LTRs (located at genomic scaffolds AE003522, from 83073 to 83278; AE003526, from 203102 to 203307; and AE003784, from 304622 to 304425), one partial insertion (genomic scaffold AE03843; from 326893 to 322718) and one complete insertion bearing inactivating mutations in the INT domain (genomic scaffold AE003787; from 212564 to 208162).

### The *Gypsy* Lineage

We have identified 24 putative new families belonging to the *Gypsy* lineage, but we were only able to obtain a consensus sequence for nine of them. Neither of the previously described Anopheles elements belonging to this lineage (*Afun1* from *A. funestus* and *Aste11* from *A. stephensi* [Cook et al. 2000]) were identified in *A. gambiae*. As shown in figure 2, we have identified an *env*-like ORF3 in three of the nine families, conforming to a R-$X_2$-R-$X_{4\text{-}5\text{-}6}$-G-$X_3$-K-$X_3$-G-$X_2$-D-$X_2$-D rule, which is slightly different from the general pattern proposed as a specific probe for the in silico detection of insect endogenous retroviral envelop protein (Terzian, Pélisson, and Bucheton 2001). The 18 insertions of families *GYPSY41_AG*, *GYPSY42_AG*, and *GYPSY43_AG*, where a target-site duplication (TSD) could be identified, showed preferential insertion at ATAT sites. Two other families of the *Gypsy* lineage (*GYPSY44_AG* and *GYPSY45_AG*) also showed

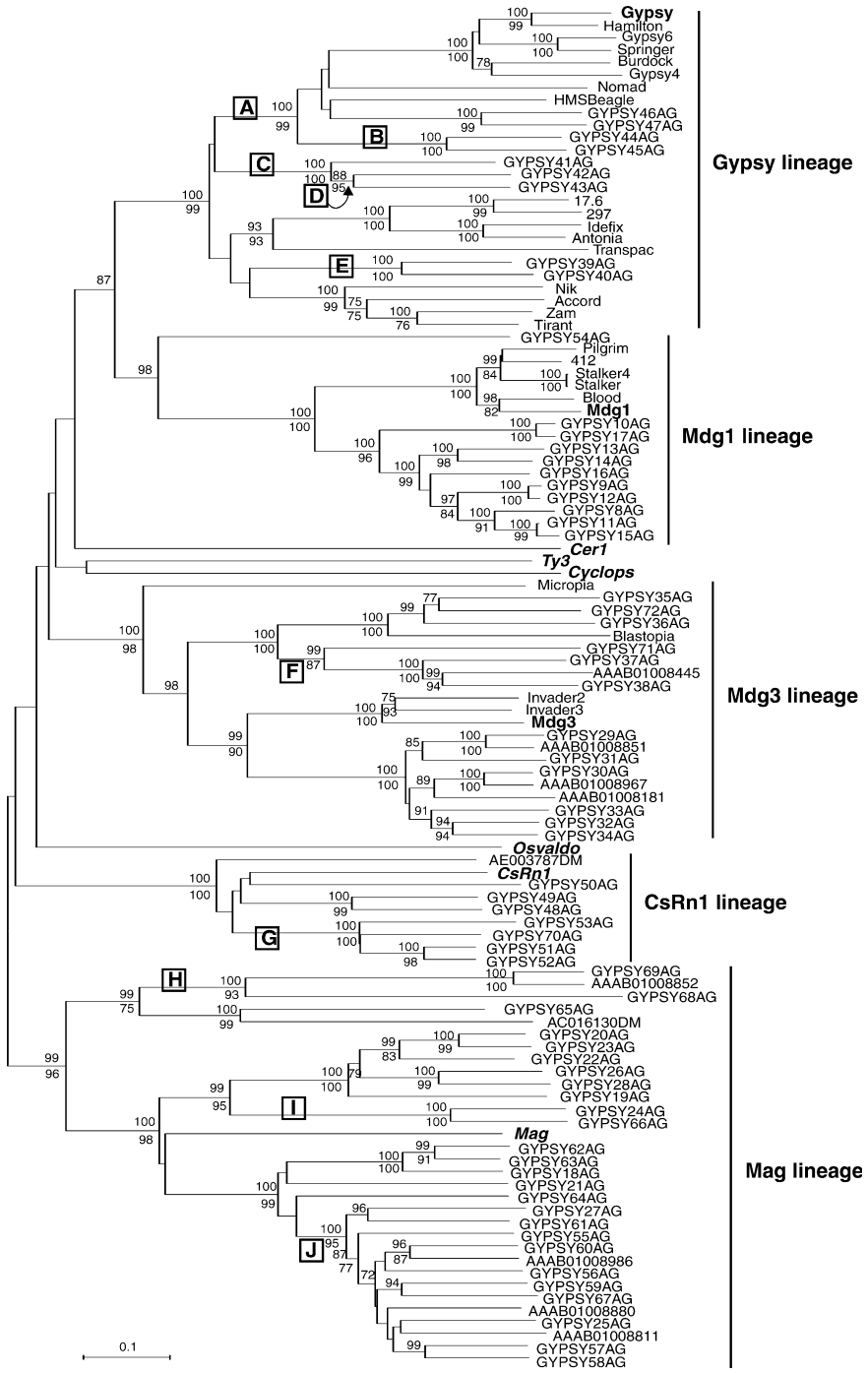



FIG. 1.—Phylogenetic relationships between the *Ty3*/*gypsy* retrotransposons of *D. melanogaster* and *A. gambiae* inferred by the NJ method based on the conserved domains of RT, RNaseH, and INT. Representative sequences of lineages not found in *D. melanogaster* are also included (represented in italics). Vertical bars indicate those lineages found in *A. gambiae*. Bootstrap values (1,000 replications) of at least 75% supporting the clusters are shown above (NJ method) or below (MP method) the branches leading to them. Specific branches (or clusters defined by these branches) referred to in the text are marked by capital letters. Names of Anopheles elements are those from Repbase (but omitting hyphens for clarity) except for putatively active individual insertions without enough information to reconstruct family consensus sequences, which are named by their accession numbers. *D. melanogaster* insertions at genomic sequences AC016130 and AE003787 are denoted by their accession numbers followed by the letters DM.

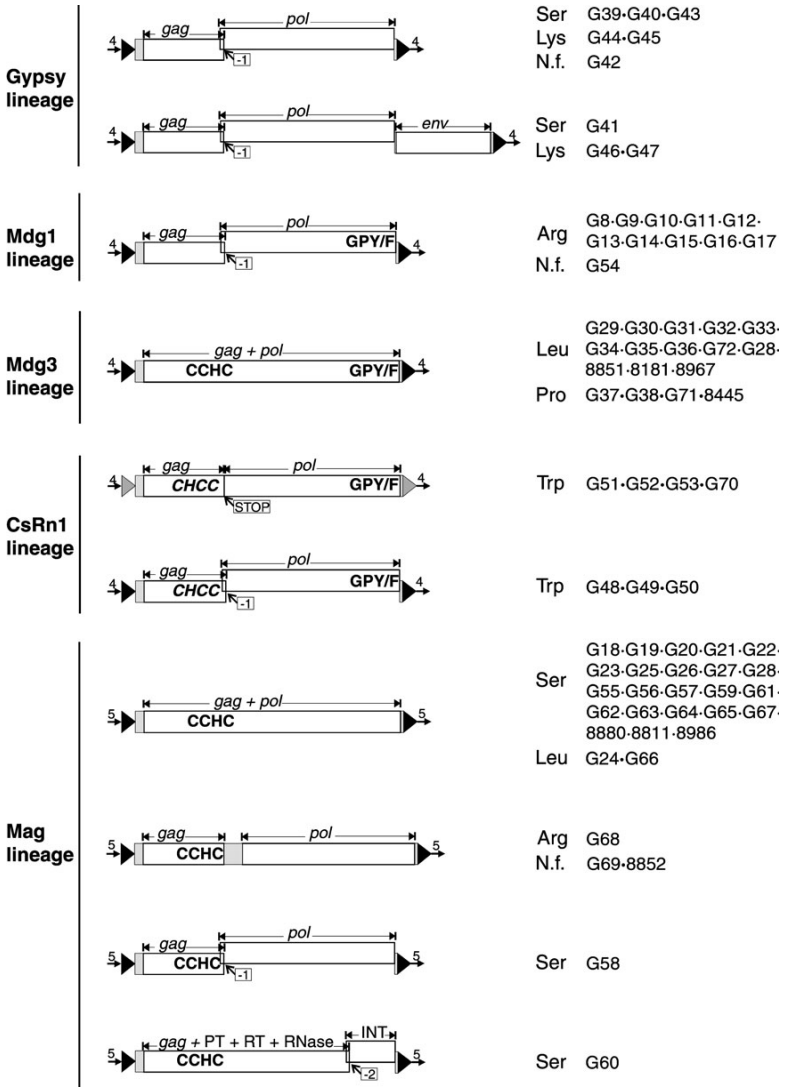



FIG. 2.—Schematic diagram showing the main structural features of the different *Ty3*/*gypsy* retrotransposons of *A. gambiae*. Open boxes indicate ORFs. Gray boxes indicate noncoding regions. Black triangles indicate LTRs with the usual TG...CA termini. Gray triangles indicate LTRs with TG...AA termini. Arrows indicate the TSD, with the number of duplicated base pairs shown above. The CCHC and GPY/F motifs are indicated (not to scale), as well as the CHCC motif presented in members of the *CsRn1* lineage instead of the usual CCHC motif (Bae et al. 2001). In all cases, the CCHC motif is present in two consecutive copies, as previously described for *Mag* from *B. mori* (Garel, Nony, and Prudhomme 1994). The tRNAs complementary to the PBS are indicated on the right side, followed by the number of families corresponding to each one of the structures and PBS. N. f. indicates that a tRNA complementary to the PBS has not been found. GYPSY has been abbreviated as G at the beginning of the Repbase family names, and the "_AG" at the end of the name has been suppressed to save space. The eight putatively active elements belonging to families without enough information to reconstruct family consensus sequences have been named according to their accession numbers, but excluding the AAAB0100 string at the beginning. Exact location of the insertions are AAAB01008880, 1172895 to 1167795; AAAB01008811, 153120 to 147785; AAAB01008986, 10714616 to 10709222; AAAB01008852, 172295 to 163857; AAAB01008851, 882036 to 887093; AAAB01008181, 18998 to 18820; AAAB01008967, 14102 to 19005; and AAAB01008445, 156 to 5229. If two ORFs overlap, a number within a box indicates the frameshift in base pairs. The word "stop" within a box indicates the *gag* stop codon that have to be read through to translate the ORF2. PT, protease; RT, reverse transcriptase; RNase, ribonuclease H; INT, integrase.

preferential insertion at C(G/T)CG, based on 12 individual members.

## The *Mag* Lineage

Two *A. gambiae* elements of the *Mag* lineage had already been partially characterized (referred to as *A. gambiae* retrotransposon 1 and *A. gambiae* retrotransposon 2 [Volff et al. 2001]). We have identified 53 putative families in the *A. gambiae* genome belonging to this lineage, but it has only been possible to characterize in detail 30 of them, representing 48% of all the characterized *Ty3*/*gypsy* families in *A. gambiae*. *A. gambiae* retrotransposon 1 and *A. gambiae* retrotransposon 2 have been

identified as members of families *GYPSY28_AG* and *GYPSY55_AG*, respectively.

So far, no *Mag*-like TEs had been identified in the genus *Drosophila*. To confirm the absence of the *Mag* lineage from *Drosophila*, we carried out Blast searches of the genome of *D. melanogaster*, using the *pol* region of different *A. gambiae* families as queries. This search led to the detection of an insertion within a 2R centromeric heterochromatin sequence (AC016130.13, unfinished sequence; *pol* region around nucleotide positions 89502 to 91566), most similar to elements from the *Mag* lineage. The insertion bears several inactivating mutations. Additional hits related to this element were identified in unfinished genomic sequences. Phylogenetic analyses revealed that this element and other *Mag* families cluster together with high bootstrap values, representing and old branch of the lineage (fig. 1).

### The *Mdg1* lineage

Most mosquito members of this lineage have been described elsewhere (Tubío, Costas, and Naveira 2004). Here, we show the existence of a basal member of this lineage (*GYPSY54_AG*). Six additional putative families related to this one had to be excluded from the analysis (table 1 in Supplementary Material online). All the other *A. gambiae* families of this lineage are more related to *D. melanogaster* families than to this novel family (fig. 1). Nevertheless, both the phylogenetic relationships (well-supported by bootstrap values in the case of NJ) and the structural characteristics of this family are consistent with its classification within the *Mdg1* lineage. Namely, the element lacks a CCHC domain, contains a GPY/F domain, and the translation of the *pol* ORF requires a frameshift of $-1$ bp as in the remaining elements of the lineage (fig. 2 [Tubío, Costas, and Naveira 2004]), although we were not able to identify any tRNA complementary to the PBS. Blast searches failed to identify elements closely related to this one in other genomes.

### The *Mdg3* Lineage

We have identified 25 putative new families in the *A. gambiae* genome belonging to this lineage, but it has only been possible to offer a full description of 16 of them. No *Mdg3* lineage elements had been described in *A. gambiae* before this work.

### Analysis of Individual Insertions

Table 1 shows the total number of insertions, classified as putatively active insertions, inactive insertions, and solo LTRs, belonging to each one of the families, as well as the chromosomal distribution of all the insertions. The most abundant family is *GYPSY50_AG*, containing 28 members. Five additional families are constituted by more than 20 members. We have identified putatively active members for 47 of the 63 characterized families. Nevertheless, it must be pointed out that at least some of the unclassified members (those meeting the criteria to be considered active but with short stretches of unfinished sequence) are most likely to be active. For 71 of the 85 putatively active elements, belonging to 40 different families, the two LTRs are identical in sequence. In addition, all families present inactive members, which have accumulated several indels and/or nonsense mutations. In general, the number of putatively active members per family is lower than that of inactive members. It was possible to calculate the average pairwise identity between putatively active copies and between inactive copies for 17 families (table 2). In all but one case, the identity was higher between putatively active copies. This difference is highly significant (Student's $t$-test $= 6.319$, $P < 0.001$). The average pairwise identity between putatively active copies was higher than 99% in the 17 families.

Assuming that most unmapped scaffolds are located in heterochromatin (as in Kaminker et al. [2002]), 89% of the putatively active elements (76/85) and 66% of the inactive elements (292/440) are inserted in euchromatin, representing a significant association between activity status and chromatin location ($\chi^2 = 18.05$, $P < 0.001$). We also checked for any biased distribution of insertions associated to particular chromosomes. The X chromosome represents 9.45% of the DNA in chromosome arms (Holt et al. 2002) but contains 12.7% of the located insertions ($\chi^2 = 5.78$, $P < 0.016$). However, this comparison does not take into account either the fact that X chromosomes are actually three quarters as numerous as any autosome in the population (*A. gambiae* males are hemizygous) or the contribution of Y-linked insertions to the pool of retrotranspositions. These two factors are conveniently addressed in the mathematical developments of the equiproportional hypothesis (Montgomery, Charlesworth, and Langley 1987; Langley et al. 1988), which produces a value of 8.0% for the expected proportion of elements on the X chromosome of *A. gambiae* (see last section of *Materials and Methods*). Observed frequencies were found to depart significantly from these expectations ($\chi^2 = 14.02$, $P < 0.001$), because an overrepresentation of insertions on the X chromosome.

## Discussion

### Diversity and Characteristic Features of the *Ty3/gypsy* Group of LTR Retrotransposons Within the *A. gambiae* Genome

In our search of the *A. gambiae* genome, we have identified retrotransposon families belonging to five of the nine major lineages of the *Ty3/gypsy* group of LTR retrotransposons. The *Osvaldo* lineage has been detected in *Drosophila* but not in mosquito. The best hits in our Blast searches using *Osvaldo* elements as query correspond to elements from other lineages. Originally discovered in *D. buzzatii* (Pantazidis, Labrador, and Fontdevila 1999), Kapitonov and Jurka (2003) have recently identified *Osvaldo*-like elements in *D. melanogaster*. Three lineages (namely *Mdg1, Gypsy*, and *Mdg3*) are represented in the genome of both *A. gambiae* and *D. melanogaster* by several distinct families. These lineages show two contrasting tree topologies. On the one hand, the *Mdg3* and *Gypsy* lineages contain several branches that comprise elements in both *Drosophila* and Anopheles, clear indication of an old diversification process before the split

**Table 1**
**Total Number of Insertions According to Its Category and Chromosomal Location**

| Family Name[a] | Total Insertions[b] | Active Insertions[c] | Inactive Insertions | Solo LTR | Chr2 | Chr3 | ChrX | NP |
|---|---|---|---|---|---|---|---|---|
| G18 | 22 | 6(6) | 10 | 3 | 9 | 8 | 2 | 3 |
| G19 | 7 | 2(2) | 5 | 0 | 2 | 0 | 0 | 5 |
| G20 | 10 | 1(1) | 5 | 3 | 3 | 2 | 1 | 4 |
| G21 | 23 | 2(2) | 12 | 7 | 13 | 7 | 2 | 1 |
| G22 | 4 | 2(1) | 2 | 0 | 0 | 2 | 0 | 2 |
| G23 | 4 | 1(0) | 1 | 1 | 0 | 2 | 1 | 1 |
| G24 | 6 | 0(0) | 4 | 1 | 2 | 1 | 0 | 3 |
| G25 | 9 | 1(1) | 5 | 2 | 5 | 2 | 0 | 2 |
| G26 | 6 | 0(0) | 4 | 0 | 2 | 2 | 0 | 2 |
| G27 | 5 | 1(0) | 2 | 1 | 1 | 1 | 1 | 2 |
| G28 | 4 | 0(0) | 1 | 2 | 2 | 0 | 1 | 1 |
| G55 | 12 | 0(0) | 3 | 6 | 2 | 6 | 3 | 1 |
| G56 | 7 | 1(1) | 3 | 1 | 2 | 0 | 2 | 3 |
| G57 | 15 | 2(2) | 5 | 6 | 4 | 7 | 1 | 3 |
| G58 | 14 | 2(1) | 2 | 7 | 8 | 4 | 0 | 2 |
| G59 | 8 | 0(0) | 4 | 2 | 3 | 1 | 2 | 2 |
| G60 | 8 | 2(2) | 3 | 0 | 2 | 5 | 0 | 1 |
| G61 | 7 | 1(1) | 2 | 2 | 5 | 0 | 0 | 2 |
| G62 | 10 | 2(2) | 4 | 3 | 5 | 4 | 1 | 0 |
| G63 | 14 | 1(1) | 8 | 4 | 5 | 4 | 0 | 5 |
| G64 | 7 | 1(1) | 3 | 1 | 3 | 2 | 0 | 2 |
| G65 | 5 | 2(2) | 3 | 0 | 0 | 3 | 0 | 2 |
| G66 | 6 | 1(0) | 4 | 0 | 0 | 2 | 0 | 4 |
| G67 | 11 | 0(0) | 6 | 3 | 2 | 3 | 2 | 4 |
| G68 | 6 | 2(1) | 3 | 0 | 4 | 0 | 1 | 1 |
| G69 | 10 | 1(1) | 5 | 1 | 2 | 4 | 3 | 1 |
| 8880 | 11 | 1(0) | 5 | 3 | 5 | 2 | 0 | 4 |
| 8811 | 7 | 1(1) | 3 | 2 | 2 | 1 | 2 | 2 |
| 8986 | 4 | 1(1) | 2 | 1 | 0 | 1 | 0 | 3 |
| 8852 | 4 | 1(1) | 1 | 2 | 0 | 0 | 1 | 3 |
| G29 | 7 | 4(4) | 2 | 1 | 3 | 1 | 2 | 1 |
| G30 | 18 | 3(2) | 3 | 6 | 5 | 6 | 1 | 6 |
| G31 | 11 | 2(1) | 2 | 6 | 3 | 4 | 2 | 2 |
| G32 | 19 | 7(7) | 4 | 2 | 6 | 10 | 1 | 2 |
| G33 | 6 | 0(0) | 3 | 3 | 0 | 5 | 0 | 1 |
| G34 | 15 | 1(1) | 7 | 3 | 4 | 3 | 1 | 7 |
| G35 | 18 | 4(3) | 4 | 3 | 11 | 6 | 0 | 1 |
| G36 | 7 | 1(1) | 2 | 2 | 4 | 2 | 0 | 1 |
| G37 | 8 | 1(1) | 3 | 3 | 1 | 7 | 0 | 0 |
| G38 | 11 | 0(0) | 4 | 2 | 6 | 3 | 1 | 1 |
| G71 | 15 | 0(0) | 5 | 8 | 1 | 5 | 1 | 8 |
| G72 | 7 | 0(0) | 4 | 1 | 2 | 1 | 1 | 3 |
| 8851 | 3 | 1(1) | 1 | 1 | 1 | 0 | 0 | 2 |
| 8181 | 5 | 1(0) | 2 | 2 | 2 | 0 | 1 | 2 |
| 8967 | 10 | 1(0) | 4 | 5 | 2 | 1 | 2 | 5 |
| 8445 | 5 | 1(1) | 3 | 0 | 2 | 2 | 0 | 1 |
| G39 | 23 | 1(1) | 6 | 11 | 6 | 7 | 3 | 7 |
| G40 | 6 | 0(0) | 3 | 2 | 2 | 1 | 2 | 1 |
| G41 | 12 | 0(0) | 6 | 6 | 6 | 2 | 2 | 2 |
| G42 | 7 | 1(0) | 2 | 1 | 0 | 1 | 2 | 4 |
| G43 | 4 | 0(0) | 4 | 0 | 1 | 0 | 0 | 3 |
| G44 | 6 | 1(1) | 3 | 1 | 3 | 0 | 0 | 3 |
| G45 | 9 | 1(1) | 2 | 4 | 2 | 0 | 2 | 5 |
| G46 | 14 | 2(1) | 5 | 6 | 6 | 5 | 1 | 2 |
| G47 | 11 | 0(0) | 7 | 4 | 3 | 1 | 2 | 5 |
| G48 | 7 | 1(1) | 2 | 2 | 1 | 2 | 0 | 4 |
| G49 | 4 | 1(1) | 2 | 0 | 0 | 2 | 1 | 1 |
| G50 | 28 | 5(5) | 14 | 7 | 15 | 10 | 1 | 2 |
| G51 | 25 | 5(5) | 6 | 3 | 14 | 6 | 1 | 4 |
| G52 | 23 | 1(1) | 14 | 6 | 7 | 6 | 1 | 9 |
| G53 | 8 | 0(0) | 5 | 2 | 2 | 2 | 0 | 4 |
| G70 | 16 | 0(0) | 8 | 4 | 4 | 6 | 0 | 6 |
| G54 | 8 | 1(1) | 2 | 5 | 2 | 2 | 2 | 2 |
| Total | 642 | 85(71) | 264 | 176 | 220 | 185 | 59 | 178 |

NOTE.—Chr2 indicates chrmosome 2, chr3 indicates chromosome 3, and chrX indicates the X chromosome. NP indicates not placed.

[a] Family names as in figure 2.

[b] The total number of insertions is the sum of putatively active sequences, inactive sequences, solo LTRs, and insertions whose activity status could not be determined because of incomplete sequencing.

[c] Active insertions with identical flanking LTRs are indicated in parenthesis.

**Table 2**
**Average Pairwise Identity Between Putatively Active Insertions and Between Inactive Insertions of the Same Family**

| Family Name[a] | $Id_{act}$[b] | Range | N[c] | $Id_{in}$[d] | Range | N[c] |
|---|---|---|---|---|---|---|
| G18 | 99.93 | 99.87–99.98 | 15 | 94.94 | 91.16–99.58 | 35 |
| G19 | 99.91 | — | 1 | 96.93 | 94.69–99.91 | 9 |
| G21 | 99.49 | — | 1 | 94.20 | 90.56–98.26 | 32 |
| G22 | 99.89 | — | 1 | 99.94 | — | 1 |
| G57 | 99.29 | — | 1 | 95.40 | 91.75–98.69 | 9 |
| G60 | 99.98 | — | 1 | 94.79 | 93.39–96.23 | 3 |
| G62 | 99.89 | — | 1 | 91.21 | 88.14–95.78 | 5 |
| G65 | 99.92 | — | 1 | 96.80 | 95.79–98.33 | 5 |
| G68[e] | 99.85 | — | 1 | 96.65 | 95.10–99.63 | 3 |
| G29 | 99.90 | 99.86–99.94 | 6 | 99.78 | — | 1 |
| G30 | 99.93 | 99.92–99.94 | 3 | 96.16 | — | 1 |
| G31 | 99.92 | — | 1 | 93.92 | — | 1 |
| G32 | 99.76 | 99.27–100 | 21 | 97.55 | 96.81–98.75 | 6 |
| G35 | 99.17 | 98.93–99.32 | 6 | 98.70 | 98.04–99.72 | 6 |
| G46 | 99.41 | — | 1 | 96.30 | 92.14–99.20 | 5 |
| G50 | 99.98 | 99.95–100 | 10 | 95.70 | 91.67–99.99 | 40 |
| G51 | 99.12 | 98.81–99.24 | 10 | 97.57 | 97.01–97.93 | 10 |

[a] Family names as in figure 2.
[b] Average pairwise identity between active copies, considering only those comparisons of at least 500 bp.
[c] Number of pairwise comparisons included in the analysis.
[d] Average pairwise identity between inactive copies, considering only those comparisons of at least 500 bp.
[e] Positions 2083 to 2371 of the consensus sequence were removed from the comparisons because of the unreliability of the alignment due to the existence of low complexity sequences.

of Diptera. On the other hand, the *Mdg1* lineage splits into two species-specific monophyletic groups, with the exception of the very basal family *GYPSY54_AG*. This fact, consistent with vertical transmission, strongly suggests that the main diversification of this lineage took place in parallel in both genomes after the divergence of flies and mosquitoes. The latest splits of Anopheles families, dated to 14 to 5 MYA (Tubío, Costas, and Naveira 2004), occurred in this lineage.

The other two mosquito lineages are almost absent from *D. melanogaster* but abundant in *A. gambiae*. The most extreme case is that of the *Mag* lineage. Thus, whereas we have identified for the first time a family of *Mag*-like elements in *D. melanogaster*, most probably consisting of old inactive members, *A. gambiae* contains at least 30 families, most of them putatively active, with an extraordinary structural diversity, accounting for approximately 48% of all the characterized *Ty3/gypsy* families in *A. gambiae* (fig. 2). Interestingly, 13 of the 30 families arose in a short period of evolutionary time (cluster J in figure 1). A similar situation occurs in the case of the *CsRn1* lineage. We have characterized seven families belonging to this lineage in *A. gambiae*, whereas the *D. melanogaster* genome bears only a single family (Bae et al. 2001) with just two full-length members bearing inactivating mutations. In summary, our data reveal a rich diversity of LTR retrotransposons in *A. gambiae*, clearly greater than in *D. melanogaster*. A similar situation has been recently described in the case of non-LTR retrotransposons (Biedler and Tu 2003).

The seven new *CsRn1* families characterized for the first time in this paper constitute a significant contribution to the total number of known families of this lineage, detected mainly in Trematoda (Bae et al. 2001; Copeland et al. 2003). This fact allows us to confirm the general distinctive characteristics of the *CsRn1* lineage, such as a PBS complementary to $tRNA^{Trp}$, the unusual CHCC gag motif instead of the typical CCHC motif, and the existence of the GPY/F motif at the 3′ end of the INT gene (Bae et al. 2001). In a similar way, each lineage shows distinctive features (or combination of features), as clearly shown in figure 2 from the *A. gambiae* representatives. Thus, members of the related lineages *Gypsy* and *Mdg1* are characterized by a frameshift of −1 bp at the *gag-pol* boundary and the absence of the CCHC motif at the C-terminal end of *gag*. They differ in the presence (*Mdg1*) or absence (*Gypsy*) of the GPY/F motif. Members of the *Mdg3* lineage are the only ones with a single *gag-pol* ORF bearing the CCHC *gag* motif and a GPY/F domain at the C-terminal end. Members of the *Mag* lineage also have the conventional CCHC *gag* motif, but they lack the GPY/F motif at the 3′ end of the INT gene. In addition, it is the unique lineage that causes a TSD at the insertion site of 5 bp, instead of the typical 4 bp.

Several characteristics, although in general are conserved within members from the same lineage, evolved in some specific branches of the phylogenetic tree. One example of this is the PBS. Thus, four closely related *Mdg3* families from *A. gambiae* (AAAB01008445, *GYPSY37_AG*, *GYPSY38_AG*, and *GYPSY71_AG*), clustered together with a strong bootstrap support (cluster F in figure 1), seem to shift from a PBS complementary to $tRNA^{Leu}$, common to all the other elements from the *Mdg3* lineage, to another complementary to $tRNA^{Pro}$. A similar situation has been detected in the case of the *Mag*-like elements *GYPSY24_AG* and *GYPSY66_AG* (cluster I in figure 1). These elements contain a PBS complementary to $tRNA^{Leu}$ instead of one complementary to $tRNA^{Ser}$ as the other members of the *Mag* lineage, with the exception of *GYPSY68_AG*, which contains a PBS complementary to $tRNA^{Arg}$. Members of the *Gypsy* lineage may be further

split into two groups, based on the presence of a PBS complementary to $tRNA^{Ser}$ or to $tRNA^{Lys}$, as pointed out previously in the case of *Drosophila* elements (Terzian, Pélisson, and Bucheton 2001). The phylogenetic tree of figure 1 strongly suggests that the PBS complementary to $tRNA^{Lys}$ arose later in a specific branch of the tree (cluster A in figure 1). Interestingly, this acquisition predated the split of Diptera.

Another clear example of distinctive characteristics evolving at specific branches is the TRS at the *gag-pol* boundary. Thus, a wide variety of strategies have been identified within the *Mag* lineage. Although most elements contain a single ORF, there is one element presenting a −1 frameshifting (characteristic of other lineages) and a cluster of three elements (cluster H in figure 1) showing two nonoverlapping ORFs separated by more than 100 bp. A similar TRS has been previously observed in several plant retrotransposons, but the mechanism to express *pol* in these cases is not clear, although splicing, internal ribosomal entry, or a bypass mechanism have been suggested (Gao et al. 2003). Furthermore, there is a mosquito family of the *Mag* lineage characterized by a long ORF encoding all the protein domains with the exception of INT, which is encoded by a different overlapping ORF, requiring a frameshift of −2 to be translated. The reasons for this particular structure are unknown.

The *CsRn1* lineage also shows different TRS. While some elements show a conventional −1 frameshifting, there is a cluster (cluster G in figure 1) that has a stop codon at the *gag-pol* boundary. Stop codon readthrough has been previously described in a few elements, such as the *Kamikaze* element from *B. mori*, the *RIRE2* element from rice and several mammalian retroviruses (revised in Gao et al. [2003]). Interestingly, the LTRs termini of members of this cluster are TG...AA, instead of the expected TG...CA. Thus, all *Drosophila* retrotransposons have the TG...CA termini except those from the *Gypsy* lineage that show AG...YT at LTR ends (Kapitonov and Jurka 2003) and the *Drosophila* family from this lineage (AE003787) that shows TG...TA (Bae et al. 2001).

We have also identified two clear cases of acquisition of preferential insertion in specific sequences. Thus, those elements belonging to the cluster formed by *GYPSY41_AG*, *GYPSY42_AG*, and *GYPSY43_AG* (cluster C in figure 1) are inserted at ATAT sites and those belonging to the related families *GYPSY44_AG* and *GYPSY45_AG* are inserted at C(G/T)CG sites (cluster B in figure 1). This preferential insertion might play an important role in host-retrotransposon coevolution (SanMiguel et al. 1996; Voytas 1996).

Finally, the *env* ORF present in some members of the *Gypsy* lineage deserves more attention. It has been shown that this lineage has acquired its *env* gene from a class of insect baculoviruses early in its evolution (Malik, Henikoff, and Eickbush 2000). Later, a few *Drosophila* elements have lost the *env* gene, such as *Burdock* (Terzian, Pélisson, and Bucheton 2001). Our survey of elements of the *Gypsy* lineage in *A. gambiae* revealed the existence of nine families. Only three of them (namely *GYPSY41_AG*, *GYPSY46_AG*, and *GYPSY47_AG*) conserve the *env* gene. Taking into account the phylogenetic relationships shown in figure 1, this fact implies three independent losses of the *env* ORF during the evolution of these elements (branches B, D, and E in figure 1). The role of the *env* gene in the life cycle of the elements from the *Gypsy* lineage remains enigmatic. It has been shown that the *env* protein of *Gypsy* may confer infectious properties to the element (Song et al. 1994), leading to the suggestion of a mechanism for *Gypsy* mobilization through infection of the germline by retroviral particles produced in the follicle cells (Song et al. 1997). Nevertheless, amplification of *Gypsy* of *D. melanogaster* may occur in an *env*-independent manner in the female germline (Chalvet et al. 1999). In a similar way, the *Drosophila Zam* element, which contains an *env* ORF, enters the oocyte via the vitelline granule traffic with no apparent need for its *env* protein, after expression in follicle cells surrounding the oocyte (Leblanc et al. 2000).

## Turnover of LTR Retrotransposons in *A. gambiae*

We have identified 47 families of the *Ty3*/*gyspy* group in mosquito containing putatively active elements (table 1). The average pairwise identity between the putatively active elements of each one of the families is always higher than 99% (table 2). Furthermore, 83.5% of the putatively active elements have identical flanking LTRs (table 1). Thus, the genome of the PEST strain of *A. gambiae*, the strain selected by the Anopheles genome project, presents clear evidence of recent activity for around 75% of the LTR retrotransposon families characterized in this work.

Lerat, Rizzon, and Biémont (2003) have recently shown that, in general, the TE families of *D. melanogaster* are characterized by a high degree of homogeneity and a lack of divergent elements. By contrast, all the families of the *Ty3*/*gypsy* group in *A. gambiae* contain a significant proportion of inactive degenerated elements, bearing indels and/or nonsense mutations and showing an average pairwise divergence significantly higher than that between active members (tables 1 and 2). Thus, there are around 40% of inactive degenerated elements within the sequenced genome of *A. gambiae* but only around 13% of active copies. Eighteen percent of the insertions (117/642) correspond to elements without obvious inactivating mutations but with unsequenced gaps, precluding their classification as either active or inactive. The significant overrepresentation of inactive elements within the unmapped scaffolds strongly indicates that heterochromatin is a shelter for these degenerate copies because of lower selection against inserted elements in heterochromatin. Bearing in mind that natural selection acts against inserted elements mainly because of insertional mutations and to chromosomal rearrangements generated by ectopic exchange between different insertions (Charlesworth and Langley 1989; Charlesworth, Sniegowski, and Stephan 1994), this lower selection is easily explained by the reduced gene density and recombination rates in heterochromatic regions. The high frequency of inactive copies is a strong indication of a slower turnover rate of *Ty3*/*gypsy* retrotransposons within the genome of *A. gambiae* than within that of *D. melanogaster*, most probably reflecting a reduced efficacy of selection against insertions of *Ty3*/*gypsy* retrotransposons in *A. gambiae*. The weakening of the efficacy of selection against TE insertions might be

related to the complex genetic population structure of *A. gambiae sensu stricto* (della Torre et al. 2002). This species is composed of different isolated or semiisolated genetic units. There are different chromosomal and molecular forms showing incomplete premating barriers. This complex structure might give rise to a reduced effective population size and/or a reduced recombination rate, both features leading to a reduced efficiency of selection against TE insertions (Charlesworth and Langley 1989; Charlesworth, Sniegowski, and Stephan 1994).

In addition to the high frequency of inactive members, we have noted a strong excess of solo LTRs in *A. gambiae* in comparison with *D. melanogaster*, confirming our previous observation from the *Mdg1* lineage (Tubío, Costas, and Naveira 2004). Thus, 176 of the 642 insertions (27%) identified in the present work correspond to solo LTRs versus 58 of 740 insertions (7.8%) reported by Kaminker et al. (2002) in *D. melanogaster*. This feature is in agreement to the above-mentioned slower turnover rate of retrotransposons in Anopheles. As a consequence, each individual insertion remains within the genome for a longer period of time, increasing the probability of exchange between the two LTRs flanking an element, giving rise to solo LTRs.

Finally, we have found a significant overrepresentation of insertions of LTR-retrotransposons on the X chromosome in comparison with the autosomes. Taking into account that *A. gambiae* exhibits comparable recombination frequencies in both sexes but males are hemizygous (Zheng et al. 1996), the overrepresentation of insertions on the X chromosome might be simply explained by a stronger selective pressure against autosomal insertions, because of their higher opportunity of ectopic recombination, according to theoretical expectations (Charlesworth and Langley 1989; Charlesworth, Sniegoski, and Stephan 1994). Nevertheless, the chromosomal distribution of TEs depends on a series of complex interacting factors in addition to recombination rates such as gene density, chromatin structure, transposition mechanisms, or interactions between TEs and host genes (Carr et al. 2002; Rizzon et al. 2002). Thus, another possibility to explain the underrepresentation of autosomal insertions might be, for instance, a lower transposition rate per copy per generation for overall male genomes. It is interesting to note that different transposition rates between sexes have been detected for specific elements (Pasyukova et al. 1997).

## Supplementary Material

Supplementary table 1.
Supplementary Alignment. Text file in fasta format. Insertions not assigned to families.

## Acknowledgments

J.M.C.T. was supported by a predoctoral fellowship from Xunta de Galicia (Spain) and FSE European funds. The authors like to thank E. Valadé for interesting discussions.

## Literature Cited

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Bae, Y-A., S-Y. Moon, Y. Kong, S-Y. Cho, and M-G. Rhyu. 2001. *CsRn1*, a novel active retrotransposon in a parasitic trematode, *Clonorchis sinensis*, discloses a new phylogenetic clade of *Ty3/gypsy*-like LTR retrotransposons. Mol. Biol. Evol. **18**:1474–1483.

Biedler, J., and Z. Tu. 2003. Non-LTR retrotransposons in the African malaria mosquito, *Anopheles gambiae*: unprecedented diversity and evidence of recent activity. Mol. Biol. Evol. **20**:1811–1825.

Boeke, J. D., T. Eickbush, S. B. Sandmeyer, and D.F. Voytas. 2000. Family *Metaviridae*. Pp. 359–367 *in* M. Regenmortel, C. Fauquet, D. Bishop, eds. Virus taxonomy: classification and nomenclature of viruses. Academic Press, San Diego.

Carr, M., J. R. Soloway, T. E. Robinson, and J. F. Brookfield. 2002. Mechanisms regulating the copy numbers of six LTR retrotransposons in the genome of *Drosophila melanogaster*. Chromosoma **110**:511–518.

Chalvet, F., L. Teysset, C. Terzian, N. Prud' homme, P. Santamaria, A. Bucheton, and A. Pelisson. 1999. Proviral amplification of the *Gypsy* endogenous retrovirus of *Drosophila melanogaster* involves *env*-independent invasion of the female germline. EMBO J. **18**:2659–2669.

Charlesworth, B., and C. H. Langley. 1989. The population genetics of *Drosophila* transposable elements. Annu. Rev. Genet. **23**:251–287.

Charlesworth, B., P. Sniegowski, and W. Stephan. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature **371**:215–220.

Cook, J. M., J. Martin, A. Lewin, R. E. Siden, and M. Tristem. 2000. Systematic screening of *Anopheles* mosquito genomes yields evidence for a mayor clade of *Pao*-like retrotransposons. Insect Mol. Biol. **9**:109–117.

Copeland, C. S., P. J. Brindley, O. Heyers, S. F. Michael, D. A. Johnston, D. L. Williams, A. C. Ivens, and B. H. Kalinna. 2003. *Boudicca*, a retrovirus-like long terminal repeat retrotransposon from the genome of the human blood fluke *Schistosoma mansoni*. J. Virol. **77**:6153–6166.

della Torre, A., C. Constantini, N. J. Besansky, A. Caccone, V. Petrarca, J. R. Powell, and M. Coluzzi. 2002. Speciation within *Anopheles gambiae*—the glass is half full. Science **298**:115–117.

Gao, X., D. J. Rowley, X. Gai, and D. F. Voytas. 2003. Translational recoding signals between *gag* and *pol* in diverse LTR retrotransposons. RNA **9**:1422–1430.

Garel, A., P. Nony, and J. C. Prudhomme. 1994. Structural features of *mag*, a *gypsy*-like retrotransposon of *Bombyx mori*, with unusual short terminal repeats. Genetica **93**:125–137.

Hall, T. A. 1999. BioEdit: a use-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. **41**:95–98.

Hill, S. R., S. S. Leung, N. L. Quercia, D. Vasiliauskas, J. Yu, I. Pasic, D. Leung, A. Tran, and P. Romans. 2001. *Ikiara* insertions reveal five new *Anopheles gambiae* transposable elements in islands of repetitious sequence. J. Mol. Evol. **52**:215–231.

Holmes, I. 2002. Transcendent elements: whole-genome transposon screens and open evolutionary questions. Genome Res. **12**:1152–1155.

Holt, R. A., G. M. Subramanian, A. Halpern et al. (126 coauthors). 2002. The genome sequence of the Malaria mosquito *Anopheles gambiae*. Science **298**:129–149.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. **16**:418–420.

Kaminker, J. S., C. M. Bergman, B. Kronmiller et al. (12 co-authors). 1995. The transposable elements of *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol. **3**:(research0084.1–0084.20).

Kapitonov, V., and J. Jurka. 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc. Natl. Acad. Sci. USA **100**:6569–6574.

Krzywinski, J., D. R. Nusskern, M. K. Kern, and N. Besansky. 2004. Isolation and characterization of Y chromosome sequences from the African malaria mosquito *Anopheles gambiae*. Genetics **166**:1291–1302.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics análisis software. Bioinformatics **17**:1244–1245.

Langley, C. H., E. Montgomery, R. Hudson, N. Kaplan, and B. Charlesworth. 1988. On the role of unequal exchange in the containment of transposable element copy number. Genet. Res. **52**:223–235.

Leblanc, P., S. Desset, F. Giorgi, A. R. Taddei, A. M Fausto, M. Mazzini, B. Dastugue, and C. Vaury. 2000. Lyfe cycle of an endogenous retrovirus, *ZAM*, in *Drosophila melanogaster*. J. Virol. **74**:10658–10669.

Lerat, E., C. Rizzon, and C. Biémont. 2003. Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. Genome Res. **13**:1889–1896

Malik, H. S., and T. H. Eickbush. 1999. Modular evolution of the integrase domain in the *Ty3/gypsy* class of LTR retrotransposons. J. Virol. **73**:5186–5190.

Malik, H. S., S. Henikoff, and T. H. Eickbush. 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res. **10**:1307–1318.

Montgomery, E., B. Charlesworth, and C. H. Langley. 1987. A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. Genet. Res. **49**:31–41.

Pantazidis, A., M. Labrador, and A. Fontdevila. 1999. The retrotransposon *Osvaldo* from *Drosophila buzzatii* displays all structural features of a functional retrovirus. Mol. Biol. Evol. **16**:909–921.

Pasyukova, E., S. Nuzhdin, W. Li, and A. J. Flavell. 1997. Germ line transposition of the copia retrotransposon in *Drosophila melanogaster* is restricted to males by tissue-specific control of copia RNA levels. Mol. Gen. Genet. **255**:115–124.

Rizzon, C., G. Marais, M. Gouy, and C. Biémont. 2002. Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. Genome Res. **12**:400–407.

SanMiguel, P., A. Tikhonov, Y-K. Jin et al. (12 co-authors). 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science **274**:765–768.

Song, S. U., T. Gerasinova, M. Kurkulos, J. D. Boeke, and V. G. Corces. 1994. An *Env*-like protein encoded by *Drosophila* retroelement: evidence that *gypsy* is an infectious retrovirus. Genes Dev. **8**:2046–2057.

Song, S. U., M. Kurkulos, J. D. Boeke, and V.G. Corces. 1997. Infection of the germ line by retroviral particles produced in the follicle ells: a possible mechanism for the mobilization of the *gypsy* retroelement of *Drosophila*. Development **124**:2789–2798.

Sprinzl, M., Vassilenko, K. S., Emmerich, J., and F. Bauer. 1999. Compilation of tRNA sequences and sequences of tRNA genes. http://www.staff.uni-bayreuth.de/~btc914/search/index.html.

Tatusova, T. A., and T. L. Madden. 1999. Blast 2 sequences—a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. **174**:247–250.

Terzian, C., A. Pélisson, and A. Bucheton. 2001. Evolution and phylogeny of insect endogenous retroviruses. BMC Evol. Biol. **1**:3.

Thompson, J. D., T. J. Gibson, K. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**:4876–4882.

Tubío, J. M. C., J. C. Costas, and H. F. Naveira. 2004. Evolution of the *mdg1* lineage of the *Ty3/gypsy* group of LTR retrotransposons in *Anopheles gambiae*. Gene **330**:123–131.

Volff, J-N., C. Körting, J. Altschmied, J. Duschl, K. Sweeney, K. Wichert, A. Froschauer, and M. Schartl. 2001. *Jule* from the fish *Xiphophorus* is the first complete vertebrate *Ty3/gypsy* retrotransposon from the *mag* family. Mol. Biol. Evol. **18**:101–111.

Voytas, D. F. 1996. Retroelements in genome organization. Science **274**:737–738.

Xiong, Y., and T. H. Eickbush. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9**:3353–3362.

Zheng, L., M. Q. Benedict, A. J. Cornel, F. H. Collins, and F. C. Kafatos. 1996. An integrated genetic map of the African human malaria vector mosquito, *Anopheles gambiae*. Genetics **143**:941–952.

Billie Swalla, Associate Editor

Accepted August 30, 2004

# III.3. Artículo 3

## Genome sequence of *Aedes aegypti*, a major *Arbovirus* vector.

*Science* (2007)

RESEARCH ARTICLE

# Genome Sequence of *Aedes aegypti*, a Major Arbovirus Vector

Vishvanath Nene,[1]* Jennifer R. Wortman,[1] Daniel Lawson,[2] Brian Haas,[1] Chinnappa Kodira,[3] Zhijian (Jake) Tu,[4] Brendan Loftus,[1]† Zhiyong Xi,[5] Karyn Megy,[2] Manfred Grabherr,[3] Quinghu Ren,[1] Evgeny M. Zdobnov,[6,7,8] Neil F. Lobo,[9] Kathryn S. Campbell,[10] Susan E. Brown,[11] Maria F. Bonaldo,[12] Jingsong Zhu,[13] Steven P. Sinkins,[14] David G. Hogenkamp,[15]‡ Paolo Amedeo,[1] Peter Arensburger,[13] Peter W. Atkinson,[13] Shelby Bidwell,[1] Jim Biedler,[4] Ewan Birney,[2] Robert V. Bruggner,[9] Javier Costas,[16] Monique R. Coy,[4] Jonathan Crabtree,[1] Matt Crawford,[3] Becky deBruyn,[9] David DeCaprio,[3] Karin Eiglmeier,[17] Eric Eisenstadt,[1] Hamza El-Dorry,[18] William M. Gelbart,[10] Suely L. Gomes,[18] Martin Hammond,[2] Linda I. Hannick,[1] James R. Hogan,[9] Michael H. Holmes,[1] David Jaffe,[3] J. Spencer Johnston,[19] Ryan C. Kennedy,[9] Hean Koo,[1] Saul Kravitz,[20] Evgenia V. Kriventseva,[6] David Kulp,[21] Kurt LaButti,[3] Eduardo Lee,[1] Song Li,[4] Diane D. Lovin,[9] Chunhong Mao,[4] Evan Mauceli,[3] Carlos F. M. Menck,[22] Jason R. Miller,[1] Philip Montgomery,[3] Akio Mori,[9] Ana L. Nascimento,[23] Horacio F. Naveira,[24] Chad Nusbaum,[3] Sinéad O'Leary,[3] Joshua Orvis,[1] Mihaela Pertea,[1]§ Hadi Quesneville,[25] Kyanne R. Reidenbach,[15] Yu-Hui Rogers,[20] Charles W. Roth,[17] Jennifer R. Schneider,[9] Michael Schatz,[1]§ Martin Shumway,[1] Mario Stanke,[26,27] Eric O. Stinson,[9] Jose M. C. Tubio,[28] Janice P. VanZee,[15] Sergio Verjovski-Almeida,[18] Doreen Werner,[27] Owen White,[1] Stefan Wyder,[6] Qiandong Zeng,[3] Qi Zhao,[1] Yongmei Zhao,[1] Catherine A. Hill,[15] Alexander S. Raikhel,[13] Marcelo B. Soares,[12] Dennis L. Knudson,[11] Norman H. Lee,[1]|| James Galagan,[3] Steven L. Salzberg,[1]§ Ian T. Paulsen,[1] George Dimopoulos,[5] Frank H. Collins,[9] Bruce Birren,[3] Claire M. Fraser-Liggett,[1]# David W. Severson[9]*

We present a draft sequence of the genome of *Aedes aegypti*, the primary vector for yellow fever and dengue fever, which at ~1376 million base pairs is about 5 times the size of the genome of the malaria vector *Anopheles gambiae*. Nearly 50% of the *Ae. aegypti* genome consists of transposable elements. These contribute to a factor of ~4 to 6 increase in average gene length and in sizes of intergenic regions relative to *An. gambiae* and *Drosophila melanogaster*. Nonetheless, chromosomal synteny is generally maintained among all three insects, although conservation of orthologous gene order is higher (by a factor of ~2) between the mosquito species than between either of them and the fruit fly. An increase in genes encoding odorant binding, cytochrome P450, and cuticle domains relative to *An. gambiae* suggests that members of these protein families underpin some of the biological differences between the two mosquito species.

Mosquitoes are vectors of many important human diseases. Transmission of arboviruses is largely associated with the subfamily Culicinae, lymphatic filarial worms with both the Culicinae and the subfamily Anophelinae, and transmission of malaria-causing parasites with the Anophelinae (*1*). *Aedes aegypti* is the best-characterized species within the Culicinae (*2*), primarily because of its easy transition from field to laboratory culture, and has provided much of the existing information on mosquito biology, physiology, genetics, and vector competence (*3*, *4*). It maintains close association with human populations and is the principal vector of the etiological agents of yellow fever and dengue fever (*5*, *6*), as well as for the recent chikungunya fever epidemics in countries in the Indian Ocean area (*7*). Despite an effective vaccine, yellow fever remains a disease burden in Africa and parts of South America, with ~200,000 cases per year resulting in ~30,000 deaths (*5*). About 2.5 billion people are at risk for dengue, with ~50 million cases per year and ~500,000 cases of dengue hemorrhagic fever, the more serious manifestation of disease. The incidence of dengue, for which mosquito management is currently the only prevention option, is on the increase (*8*). Thus, there is an urgent need to improve the control of these diseases and their vector.

The availability of a draft sequence of the ~278 million base pair (Mbp) genome of *Anopheles gambiae* (*9*) has accelerated research to develop new mosquito- and malaria-control strategies. Comparisons between *An. gambiae* and *Drosophila melanogaster* (*10*) revealed genomic differences between the two insects that reflect their divergence ~250 million years ago (*11*). *Anopheles* mosquitoes radiated from the *Aedes* and *Culex* lineages ~150 million years ago (*12*), and *Ae. aegypti* and *An. gambiae* share similar characteristics such as anthropophily, but they exhibit variation in morphology and physiology, mating behavior, oviposition preferences, dispersal, and biting cycle (*1*). Both mosquito species have three pairs of chromosomes, but *Ae. aegypti* lacks heteromorphic sex chromosomes (*13*). To provide genomics platforms for research into *Ae. aegypti* and to harness the power of comparative genome analyses, we undertook a project to sequence the genome of this mosquito species.

**Assembly of a draft genome sequence of *Aedes aegypti*.** Whole-genome shotgun sequencing was performed on DNA purified from newly hatched larvae of an inbred substrain ($LVP^{ib12}$) of the Liverpool strain of *Ae. aegypti*, which is tolerant to inbreeding while maintaining relevant phenotypes (*14*). About 98% of the sequence, assembled using Arachne (*15*), is contained within 1257 scaffolds with an N50 scaffold size of ~1.5 Mbp (i.e., half of the assembly resides in scaffolds this size or longer). Assembly statistics for the 1376-Mbp genome are given in table S1. Data related to the genome project have been deposited in GenBank (project accession number AAGE00000000).

The genome size of *Ae. aegypti* as determined by sequence analysis is larger than the

[1]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [2]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [3]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02141, USA. [4]Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA. [5]Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA. [6]University of Geneva Medical School, 1 rue Michel-Servet, Geneva 1211, Switzerland. [7]Swiss Institute of Bioinformatics, 1 rue Michel-Servet, Geneva 1211, Switzerland. [8]Imperial College London, South Kensington Campus, London SW7 2AZ, UK. [9]University of Notre Dame, Notre Dame, IN 46556, USA. [10]Harvard University, Cambridge, MA 02138, USA. [11]College of Agricultural Sciences, Colorado State University, Fort Collins, CO 80523, USA. [12]Northwestern University, Chicago, IL 60614, USA. [13]University of California, Riverside, CA 92521, USA. [14]University of Oxford, Oxford OX1 3PS, UK. [15]Purdue University, West Lafayette, IN 47907, USA. [16]Centro Nacional de Genotipado, Fundación Pública Galega de Medicina Xenómica, Hospital Clínico Universitario de Santiago, Edif. Consultas Planta-2, Santiago de Compostela E-15706, Spain. [17]Institut Pasteur, Paris 75724, France. [18]Universidade de Sao Paulo, Instituto de Quimica, Sao Paulo SP 05508-900, Brazil. [19]Texas A&M University, College Station, TX 77843, USA. [20]Joint Technology Center, 5 Research Place, Rockville, MD 20850, USA. [21]University of Massachusetts, Amherst, MA 01003, USA. [22]Universidade de Sao Paulo, Institute of Biomedical Sciences, Sao Paulo SP 05508-900, Brazil. [23]Instituto Butantan, Sao Paulo SP 05503-900, Brazil. [24]Universidade da Coruña, A Coruña 15001, Spain. [25]Institut Jacques Monod, CNRS, Université Paris Diderot et Université Pierre-et-Marie Curie 2, Place Jussieu, Paris 75252, France. [26]507A Engineering 2, University of California, 1156 High Street, Santa Cruz, CA 95064, USA. [27]Universität Göttingen, Goldschmidtstraße 1, Göttingen 37077, Germany. [28]Complexo Hospitalario Universitario de Santiago, Santiago de Compostela 15706, Spain.

*To whom correspondence should be addressed. E-mail: nene@tigr.org (V.N.); severson.1@nd.edu (D.W.S.)
†Present address: University College Dublin, Dublin 4, Ireland.
‡Deceased.
§Present address: 3125 Biomolecular Sciences Building, University of Maryland, College Park, MD 20742, USA.
||Present address: George Washington University Medical Center, Ross Hall, Room 603, 2300 I Street, NW, Washington, DC 20037, USA.
#Present address: Institute of Genome Sciences and Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA.

original estimate, ~813 Mbp, which was based on $C_0t$ (DNA reassociation kinetics) analysis carried out in 1991 (*16*). An overinflated genome size could arise from assembled sequence data as a result of allelic sequence polymorphism present in a heterogeneous population of mosquitoes being sequenced. Although the estimate of 1376 Mbp may contain some such regions, we do not believe that our estimate is out of range by a large margin for the following reasons: (i) The strain that was used for the sequencing project was highly inbred (*14*); (ii) assembled sequences that are potentially "undercollapsed" are <5% of the estimated genome size (fig. S1); and (iii) flow cytometry data from six isolates of *Ae. aegypti*, including the parent of $LVP^{ib12}$, indicate estimated genome sizes of 1213 to 1369 Mbp (table S2).

Genetic and physical mapping data allowed assignment, but without order or orientation, of 63, 48, 39, 43, and 45 scaffolds to *Ae. aegypti* chromosome 1 and chromosome arms 2p, 2q, 3p, and 3q, respectively (*14*). These scaffolds total ~430 Mbp in length and represent ~31% of the genome (table S3). Thus, the development of high-resolution physical mapping techniques and the generation of additional random or targeted sequence data are priorities for improving the quality of the current fragmented genome assembly and size estimate. Such progress would enable unambiguous differentiation between regions of segmental duplications and residual haplotype polymorphism.

**The genome of *Aedes aegypti* is riddled with transposable elements.** Transposable elements (TEs) contribute substantially to the factor of ~5 size difference between the *Ae. aegypti* and *An. gambiae* genomes. About 47% of the *Ae. aegypti* genome consists of TEs (Fig. 1 and table S4; see table S4 legend for definitions of TE family, element, and copy). *Aedes aegypti* harbors all known types of TEs that have been reported in *An. gambiae* with the exception of two DNA transposons, *merlin* (*17*) and *gambol* (*18*). Simple and tandem repeats occupy ~6% of the genome, and an additional ~15% consists of repetitive sequences that remain to be classified.

Most eukaryotic TE families characterized to date (*19*) are present in *Ae. aegypti* and more than 1000 TEs have been annotated, representing a diverse collection of TEs in a single genome (table S4). Although the majority of protein-coding TEs appear to be degenerate, more than 200 elements have at least one copy with an intact open reading frame (ORF) and other features suggesting recent transposition. About 3% of the genome is composed of ~13,000 copies of the element *Juan-A* in the Jockey family of non–long terminal repeat (LTR) retrotransposons. A tRNA-related short interspersed nuclear element, *Feilai-B*, has the highest copy number, with ~50,000 copies per haploid genome. Only one highly degenerate *mariner* element is found in *Ae. aegypti*, whereas at least 20 *mariner* elements, many with intact ORFs, were found in *An. gambiae*. TEs present in *Ae. aegypti* but missing from *An. gambiae* include the *LOA* family of non-LTR retrotransposons, the *Osvaldo* element of the *Ty3/gypsy* LTR retrotransposons (*20*), and a unique family, *Penelope* (*21*). Comparison of *Ae. aegypti* and *An. gambiae* TE sequences is consistent with the interpretation of an overall lack of apparent horizontal transfer events, as a single candidate for such events was identified (*14*); one full-length copy of the *ITmD37E* DNA transposon in *Ae. aegypti* is 93% identical at the nucleotide level to a similarly classified TE in *An. gambiae*.

Miniature inverted repeat transposable elements (MITEs) and MITE-like elements of non–protein-coding TEs in *Ae. aegypti* have terminal inverted repeat sequences and target-site duplications, features characteristic of transposition of DNA transposons. Such TEs can be mobilized to transpose in trans, by transposases encoded by DNA transposons (*22*). The latter TEs occupy only 3% of the *Ae. aegypti* genome and they are less numerous than non–protein-coding DNA elements, which occupy 16% of the genome (table S4). Thus, DNA transposons may have contributed to the expansion in size and organization of the *Ae. aegypti* genome through cross-mobilization of MITEs and MITE-like TEs.

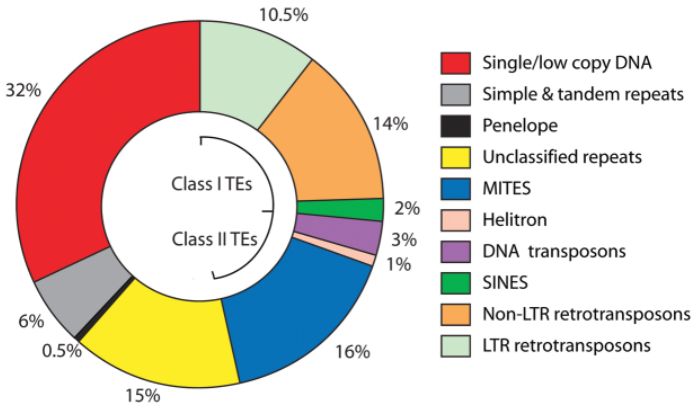



**Fig. 1.** Relative genomic content of annotated TEs and other sequences in *Aedes aegypti*. TEs have been deposited in TEfam, a relational database for submission, retrieval, and analysis of TEs (http://tefam.biochem.vt.edu).

**Annotation of the draft genome sequence.** The fragmented nature of the assembled genome sequence, an asymmetric distribution of intron lengths within genes (figs. S2 and S3), and the frequent occurrence of TE-associated ORFs close to genes and within introns complicated the process of automated gene modeling and often led to prediction of split or chimeric gene models. Thus, we developed a multistage genome masking strategy to minimize the negative effects of TEs and other repetitive elements before gene finding (resulting in masking ~70% of the genome sequence). We also optimized gene-finding programs via iterative manual inspection of predicted gene models relative to a training set (*14*).

Two independent automated pipelines for structural annotation resulted in the prediction of 17,776 and 27,284 gene models, respectively (*14*). We made extensive use of a large collection of ~265,000 *Ae. aegypti* expressed sequence tags (ESTs) and dipteran protein and cDNA sequences in producing and then merging the two data sets into a single high-confidence gene set, which consists of 15,419 gene models (AaegL1.1). Alternative splice forms derived from these genes are predicted to generate at least 16,789 transcripts. Table 1 lists some of the genome and protein-coding characteristics of *Ae. aegypti* and those of *D. melanogaster* and *An. gambiae*.

Gene descriptions and molecular function Gene Ontology (GO) codes were assigned computationally to predicted protein sequences by means of BLASTP comparison searches with protein databases (*14*). The functional annotation pipeline included analyses of protein domains as well as secretion signal sequence and transmembrane motifs. A total of 8332 proteins were assigned a description, 9335 proteins were assigned GO terms, 2796 were assigned as "hypothetical proteins," and 5027 were denoted "conserved hypothetical proteins."

Genes encoding proteins <50 amino acid residues in length were not included in this annotation release unless they encoded known small proteins. However, these and other genes are captured in a set of 15,396 lower-confidence gene models that is available for analysis as a supplementary release (*14*). On the basis of transcriptional mapping data and limited manual examination, we anticipate that ~5 to 10% of the second-tier models or modified versions of them represent "real" genes.

**TEs contribute to complex protein-coding gene structures in *Aedes aegypti*.** A striking feature of protein-coding genes in *Ae. aegypti* is the factor of 4 to 6 increase in the average

RESEARCH ARTICLE

length of a gene relative to *An. gambiae* and *D. melanogaster*, which is due to longer intron lengths rather than longer exons or an increased number of introns (Table 1). The increased length of introns is primarily due to infiltration by TEs; a plot of intron size before and after masking repeat sequences reveals a shift to shorter intron lengths (fig. S2). A more global perspective of the genome expansion was revealed by the difference in genomic span (factor of ~4.6) of conserved gene arrangements between *Ae. aegypti* and *An. gambiae* that occupy ~33% of each genome (table S5 and fig. S3), providing evidence that TE-mediated expansion in both genic and intergenic regions has contributed to the increased size of the *Ae. aegypti* genome. Long introns, in particular those in 5′ and 3′ untranslated regions, are likely to complicate in silico studies to define cis-acting transcription and translational regulatory elements, as they may be distant from coding sequences (fig. S4).

**Transcriptional analyses.** Data derived from three different transcript-profiling platforms—ESTs, massively parallel signature sequencing (MPSS), and 60-nucleotide oligomer–based microarrays—were used to experimentally confirm predicted protein-coding gene models and to gain insight into differential transcription profiles (*14*). In total, the platforms identified transcripts from 12,350 (80%) of 15,419 genes. Mapping of ~265,000 ESTs and cDNA sequences and MPSS signature sequence tags to the genome sequence as well as gene models provided evidence for transcription of 9270 and 3984 genes, respectively, whereas microarray data identified transcripts from 9143 genes (table S6). The smaller number of genes identified by MPSS (table S7) may in part be explained by the observation that only about two-thirds of the genes can be assayed by MPSS, as this approach required the presence of a Dpn II restriction enzyme site within the transcribed region. The platforms identified a common set of 2558 genes and each platform identified a unique set of genes (fig. S5), which highlights the importance of using a multi-platform approach. The data provide empirical support for ~76% of genes annotated as hypothetical (table S8), underscoring the validity of ab initio gene-finding programs in identifying novel genes.

Differences in transcript abundance between pools of RNA from nonadult developmental stages and from 4-day-old, non–blood-fed adult females were revealed by the microarray analyses, which identified 398 and 208 preadult stage and adult female enriched transcripts, respectively (table S9). Functional categorization of these transcripts differed mainly with regard to cytoskeletal, structural, and chemosensory functions (Fig. 2). Differential transcription of genes thought to be involved in chemosensory processes between these stages was conspicuous, with 17 transcripts highly enriched in mosquito developmental stages and only 3 enriched in adult females. A larger number of immune-system gene transcripts were also enriched in preadult stages (38 preadult versus 19 adult), which may reflect a broader microbial exposure of larvae and pupae in their aqueous environments. In addition, highly expressed genes encoding cuticle proteins in preadult stages (38 preadult versus 1 adult) are indicative of their function in cuticle metabolism and in a variety of other processes, including immunity, that are particularly dominant during development. The non–blood-fed status of the female mosquitoes did not enable discrimination of genes that carry out female-specific functions that mostly relate to blood processing and egg production.

***Aedes aegypti* gene families and domain composition.** Consistent with evolutionary distance estimates (*12*), there is a higher degree of similarity between the *Ae. aegypti* and *An. gambiae* proteomes than between the mosquito and *D. melanogaster* proteomes. Orthologous proteins were computed among the three genomes, with 67% of the *Ae. aegypti* proteins having an ortholog in *An. gambiae* and 58% having an ortholog in *D. melanogaster* (Fig. 3A). Analysis of three-way, single-copy orthologs revealed average amino acid identity of 74% between the mosquito proteins, in contrast with ~58% identity between mosquito and fruit fly proteins (fig. S7). About 2000 orthologs are shared only between the mosquitoes and may represent functions central to mosquito biology. Although most of these proteins are of unknown function, ~250 can be assigned a predicted function, of which 28% are involved in gustatory or olfactory systems, 12% are members of the cuticular gene family, and 8% are members of the cytochrome P450 family (*14*).

Mapping of protein domains with Interpro (*23*) revealed an expansion of zinc fingers, insect cuticle, chitin-binding peritrophin-A, cytochrome P450, odorant binding protein (OBP) A10/OS-D, and insect allergen–related domains, among others, in *Ae. aegypti* relative to *An. gambiae*, *D. melanogaster*, and the honey bee *Apis mellifera* (table S10). Some of these constitute large *Ae. aegypti* gene families, as revealed by two independent clustering methods (*14*) (table S11). Genes containing zinc finger–like domains could be of transposon or retroviral origin, and these remain to be assessed.

Species-specific differences in the number of members within a multigene family often provide clues about biological adaptation to environmental challenges. In this context, cuticle proteins have been described to play diverse roles in exoskeleton formation and wound healing and are expressed in hemocytes, a major cell type that mediates innate immunity (*24*). Cuticular proteins also are implicated in arbovirus transmission (*25*). Expansion of olfactory receptors and OBPs in *Ae. aegypti* may contribute to an elaborate olfactory system, which in turn may be linked to the expansion in detoxification capacity. The latter and insect allergen–related genes, suggested to have a digestive function, may contribute to the relative robustness of *Ae. aegypti* and also could manifest in a higher insecticide resistance. In this context, the genome and EST data have led to the development of a specific microarray to identify candidate genes among members of multigene families (cytochrome P450, glutathione *S*-transferase, and carboxylesterase) associated with metabolic resistance to insecticides (*26*). This platform will provide a means to rapidly survey mechanisms of insecticide resistance in mosquito populations and represents an important tool in managing insecticide deployment and development programs.

G protein–coupled receptors (GPCRs) that are expected to function in signal transduction cascades in *Ae. aegypti* have been manually identified (*14*). This superfamily of proteins includes 111 nonsensory class A, B, and C GPCRs, 14 atypical class D GPCRs, and 10 opsin photoreceptors (tables S12 and S13). *Aedes aegypti* possesses orthologs for >85% of the *An. gambiae* and *D. melanogaster* nonsensory GPCRs, which suggests conservation of GPCR-mediated neurological processes across the Diptera. Many *Ae. aegypti* GPCRs have sequence similarity to known drug targets (*27*) and may reveal new

**Table 1.** Comparative statistics of *Ae. aegypti* nuclear genome coding characteristics.

| Feature | Species | | |
|---|---|---|---|
| | *Ae. aegypti* | *An. gambiae*‡ | *D. melanogaster*‡ |
| Size (Mbp) | 1,376 | 272.9 | 118 |
| Number of chromosomes | 3 | 3 | 4 |
| Total G+C composition (%) | 38.2 | 40.9 | 42.5 |
| Number of protein-coding genes | 15,419 | 13,111 | 13,718 |
| Average gene length* (bp) | 14,587 | 5,124 | 3,460 |
| Average protein-coding gene length† (bp) | 1,397 | 1,154 | 1,693 |
| Percent genes with introns | 90.1 | 93.6 | 86.2 |
| Average number of exons/gene | 4.0 | 3.9 | 4.9 |
| Average intron length (bp) | 4,685 | 808 | 1,175 |
| Longest intron (bp) | 329,294 | 87,786 | 132,737 |
| Average length of intergenic region (bp) | 56,417 | 17,265 | 6,043 |

*Includes introns but not untranslated regions. †Does not include introns. ‡Statistics were derived from genome updates for *An. gambiae* R-AgamP3 and *D. melanogaster* R-4.2.

opportunities for the development of novel insecticides.

**Metabolic potential and membrane transporters.** *Aedes aegypti* and *An. gambiae* are predicted to contain similar metabolic profiles as judged by assigning an Enzyme Commission (EC) number to both mosquito proteomes (table S14). Given the early stages of annotation, it is premature to draw conclusions from missing enzymes in predicted *Ae. aegypti* metabolic pathways. For example, assignment of EC numbers to the supplemental *Ae. aegypti* gene set (table S14) resulted in the identification of an additional 12 EC numbers (table S15) not present in AaegL1.1.

An automated pipeline (*28*) was used to predict potential membrane transporters for *Ae. aegypti* and *An. gambiae*, and their transport capacity resembles that of *D. melanogaster* (table S16). Similar to other multicellular eukaryotes, ~32% of all three insect transporters code for ion channels and probably function to maintain hemolymph homeostasis under different environmental conditions by modulating the concentrations of $Na^+$, $K^+$, and $Cl^-$ ions. *Aedes aegypti* encodes 52 more paralogs of voltage-gated potassium ion channels, epithelial sodium channels, and ligand-gated ion channels than *An. gambiae* and 65 more such paralogs than *D. melanogaster*. These channels play important roles in the signal transduction pathway and cell communication in the central nervous system and at neuromuscular junctions. A collection of 64 putative adenosine triphosphate–binding cassette transporters was identified, including subgroups that encode multidrug efflux proteins. *Aedes aegypti* encodes 16 more members of four different types of amino acid transporters than *An. gambiae* and 13 more members than *D. melanogaster*. Mosquito larvae cannot synthesize de novo all the basic, neutral, or aromatic L–amino acids (*3*) and must rely on uptake of these essential amino acids. The richer repertoire of membrane transport systems in *Ae. aegypti* is likely to intersect with the apparent increase in odorant reception and detoxification capacity.

**Autosomal sex determination and sex-specific gene expression.** Heteromorphic sex chromosomes are absent in *Ae. aegypti* and other culicine mosquitoes (*13*). Instead, sex is controlled by an autosomal locus wherein the male-determining allele, *M*, is dominant. The primary switch mechanism at the top of the mosquito sex determination cascade is different from that of *D. melanogaster*, where the X-chromosome/autosome ratio controls sex differentiation. However, we expect conservation of function in mosquito orthologs of *Drosophila* genes that are further downstream of the cascade (*29*). We verified the presence of a number of these in *Ae. aegypti*, including orthologs for *doublesex*, *transformer-2*, *fruitless*, *dissatisfaction*, and *intersex* (table S17).

To define gene expression differences between the sexes, we analyzed microarray transcription profiles of 4-day-old, non–blood-fed adult female and male mosquitoes (Fig. 2); 669 and 635 transcripts were enriched in females and males, respectively, and 6713 transcripts were expressed at similar levels in both sexes (table S18). An additional 373 and 534 transcripts generated exclusive hybridization signals (with signal intensity below the cutoff threshold level in one channel) in females and males, respectively, and may therefore represent sex-specific transcripts. Functional categorization of female and male enriched transcripts yielded similar results, with some exceptions; male mosquitoes expressed a larger number of immune system–related transcripts (40 in males versus 25 in females) and redox- or stress-related transcripts (45 in males versus 33 in females). By comparing the *Ae. aegypti* profiles with previously described *An. gambiae* sex-specific microarray analyses (*30*), we identified 144 orthologous genes displaying the same sex-specific transcription pattern in *An. gambiae* (table S19), whereas 74 orthologs showed an opposite profile (table S20), suggesting differences in certain sex-specific functions between the two mosquito species.

**Conserved synteny with *Anopheles gambiae* and *Drosophila melanogaster*.** The assignment of 238 *Ae. aegypti* scaffolds containing ~5000 genes—about one-third of the predicted gene set—to a chromosomal location on the basis of genetic and physical mapping data (*14*) allowed us to compare ortholog position and to identify

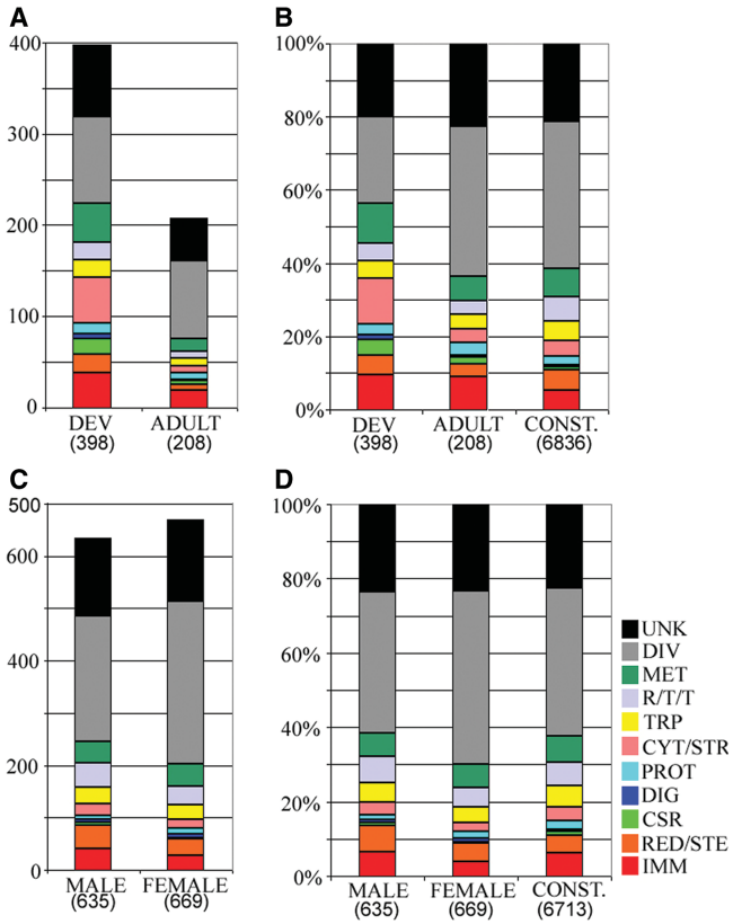



**Fig. 2.** Transcriptome analyses of *Aedes aegypti*. (**A**) Functional class distributions of genes that are enriched in preadult stages (DEV) and the adult female stage (ADULT) (table S9). (**B**) Proportions of functional gene classes, expressed as percentage of the total number of genes that are enriched in preadult stages (DEV), adult female stage (ADULT), and constitutively expressed genes (CONST.). (**C** and **D**) same as (A) and (B) for genes enriched in the male, in the female, and common (CONST.) for both sexes (table S18). Functional classes: IMM, immunity; RED/STE, redox and oxidoreductive stress; CSR, chemosensory reception; DIG, blood and sugar food digestive; PROT, proteolysis; CYT/STR, cytoskeletal and structural; TRP, transport; R/T/T, replication, transcription, and translation; MET, metabolism; DIV, diverse functions; UNK, unknown functions. The total number of genes in each category is indicated in parentheses.

RESEARCH ARTICLE

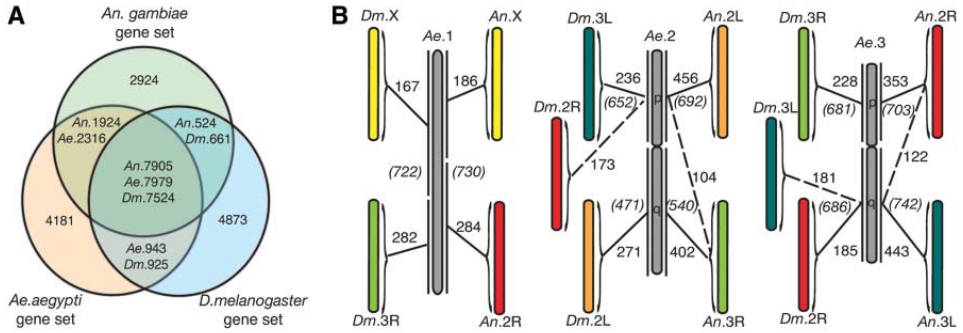



**Fig. 3.** Orthology and chromosomal synteny among *Ae. aegypti*, *An. gambiae*, and *D. melanogaster*. (**A**) Each circle represents a gene set for *Ae. aegypti* (*Ae*), *An. gambiae* (*An*), and *D. melanogaster* (*Dm*). Because a gene can be involved in several homologies, gene sets do not always have the same number of genes within intersections (e.g., in the *Ae-Dm* comparison, 943 *Ae* genes are similar to *Dm* and 925 *Dm* genes are similar to *Ae*). (**B**) *Aedes aegypti* chromosomes are represented in gray (not to scale). Chromosome arms are designated as "p" and "q"; chromosome 1 has no arm distinctions. Colored chromosomes represent the syntenic chromosome from *An. gambiae* or *D. melanogaster* (not to scale). Solid and dashed lines link each *Ae. aegypti* chromosome to its primary and secondary syntenic chromosome, respectively. The number of *Ae* orthologs to *An* and *Dm* chromosome arms is indicated, and the total number of orthologs on the *Ae* chromosome arm to *Ae* or *Dm* is shown in italics and parentheses.

conserved evolutionary associations between *Ae. aegypti* and *An. gambiae* or *D. melanogaster* chromosomes (tables S3 and S21). Most of the *Ae. aegypti* chromosome arms, with the exception of 2p and 3q, exhibited a distinct one-to-one correlation with *An. gambiae* and *D. melanogaster* chromosome arms with respect to the proportion of orthologous genes conserved between chromosome arm pairs (Fig. 3B). These findings confirm and extend previous results that compared a small number (~75) of *Ae. aegypti* genes with orthologs in *An. gambiae* and *D. melanogaster* (*31*).

Maps of conserved local gene arrangements (microsynteny) were computed by identifying blocks of at least two neighboring single-copy orthologs in each pair of genomes and allowing not more than two intervening genes (*14*). In line with the species divergence times, twice as many orthologs are similarly arranged between these mosquito species than between either of them and the fruit fly (table S22) (*32*); 1345 microsyntenic blocks were identified between *Ae. aegypti* and *An. gambiae*, containing 5265 out of a total of 6790 single-copy orthologs (tables S5 and S22). When *D. melanogaster* is used as an outgroup to count synteny breaks that have occurred in each mosquito lineage since their radiation, the data indicate a rate of genome shuffling in the *Ae. aegypti* lineage greater by a factor of ~2.5 than that in the *An. gambiae* lineage (*14*). However, this estimate may be inflated because of the fragmented nature of the current *Ae. aegypti* genome assembly. Thus, the highly repetitive nature of the *Ae. aegypti* genome appears to have facilitated local gene rearrangements, but it does not appear to have had a gross influence on chromosomal synteny.

**Concluding remarks.** The draft genome sequence of *Ae. aegypti* will stimulate efforts to elucidate interactions at the molecular level between mosquitoes and the pathogens they transmit. This already can be seen in, for example, analysis of components of the Toll immune signaling pathway (*33*) and identification of genes encoding insulin-like hormone peptides (*34*).

We expect that the sequence data will facilitate the identification of *Ae. aegypti* genes encoding recently described midgut receptors for dengue virus (*35*). Dengue vector competence is a quantitative trait, and multiple loci determine virus midgut infection and escape barriers (*36*). Unfortunately, the fragmented nature of the genome sequence and its low gene density have precluded its use in the identification of a comprehensive list of candidate genes for vector competence phenotypes or sex determination. The sequence may be used to improve the resolution of the current genetic map (*37*) and to integrate transcriptional profiling data with genetic studies (*38*), but filling gaps in the assembled sequence remains a high priority, especially when exploring genetic variations between the sequenced strain and field populations of *Ae. aegypti*.

The ongoing genome project on *Culex pipiens quinquefasciatus*, a vector for lymphatic filariasis and West Nile virus, will provide additional resources to underpin studies to systematically study common and mosquito species–specific gene function. Such analyses should improve our understanding of mosquito biology and the complex role of mosquitoes in the transmission of pathogens, and may result in the development of new approaches for vector-targeted control of disease.

### References and Notes


1. B. J. Beaty, W. C. Marquardt, *Biology of Disease Vectors* (Univ. Press of Colorado, Niwot, CO, ed. 1, 1996).
2. S. R. Christophers, Aedes aegypti *(L.): The Yellow Fever Mosquito, Its Life History, Bionomics and Structure* (Cambridge Univ. Press, Cambridge, 1960).
3. A. N. Clements, *The Biology of Mosquitoes* (Chapman & Hall, London, 1992).
4. D. W. Severson, S. E. Brown, D. L. Knudson, *Annu. Rev. Entomol.* **46**, 183 (2001).
5. O. Tomori, *Crit. Rev. Clin. Lab. Sci.* **41**, 391 (2004).
6. World Health Organization, *Dengue and Dengue Haemorrhagic Fever* (World Health Organization, Geneva, 2002).
7. B. L. Ligon, *Semin. Pediatr. Infect. Dis.* **17**, 99 (2006).
8. J. S. Mackenzie, D. J. Gubler, L. R. Petersen, *Nat. Med.* **10**, S98 (2004).
9. R. A. Holt *et al.*, *Science* **298**, 129 (2002).
10. E. M. Zdobnov *et al.*, *Science* **298**, 149 (2002).
11. M. W. Gaunt, M. A. Miles, *Mol. Biol. Evol.* **19**, 748 (2002).
12. J. Krzywinski, O. G. Grushko, N. J. Besansky, *Mol. Phylogenet. Evol.* **39**, 417 (2006).
13. G. B. J. Craig, W. A. Hickey, in *Genetics of Insect Vectors of Disease*, J. W. Wright, R. Pal, Eds. (Elsevier, New York, 1967), pp. 67–131.
14. See supporting material on *Science* Online.
15. D. B. Jaffe *et al.*, *Genome Res.* **13**, 91 (2003).
16. A. M. Warren, J. M. Crampton, *Genet. Res.* **58**, 225 (1991).
17. C. Feschotte, *Mol. Biol. Evol.* **21**, 1769 (2004).
18. M. R. Coy, Z. Tu, *Insect Mol. Biol.* **14**, 537 (2005).
19. N. Craig, R. Cragie, M. Gellert, A. Lambowitz, Eds., *Mobile DNA II* (American Society for Microbiology Press, Washington, DC, 2002).
20. J. M. Tubio, H. Naveira, J. Costas, *Mol. Biol. Evol.* **22**, 29 (2005).
21. I. R. Arkhipova, K. I. Pyatkov, M. Meselson, M. B. Evgen'ev, *Nat. Genet.* **33**, 123 (2003).
22. X. Zhang, N. Jiang, C. Feschotte, S. R. Wessler, *Genetics* **166**, 971 (2004).
23. E. M. Zdobnov, R. Apweiler, *Bioinformatics* **17**, 847 (2001).
24. L. C. Bartholomay *et al.*, *Infect. Immun.* **72**, 4114 (2004).
25. H. R. Sanders *et al.*, *Insect Biochem. Mol. Biol.* **35**, 1293 (2005).
26. H. Ranson, personal communication.
27. A. Wise, K. Gearing, S. Rees, *Drug Discov. Today* **7**, 235 (2002).
28. Q. Ren, K. H. Kang, I. T. Paulsen, *Nucleic Acids Res.* **32**, D284 (2004).
29. C. Schutt, R. Nothiger, *Development* **127**, 667 (2000).
30. O. Marinotti *et al.*, *Insect Mol. Biol.* **15**, 1 (2006).
31. D. W. Severson *et al.*, *J. Hered.* **95**, 103 (2004).
32. E. M. Zdobnov, P. Bork, *Trends Genet.* **23**, 16 (2007).
33. S. W. Shin, G. Bian, A. S. Raikhel, *J. Biol. Chem.* **281**, 39388 (2006).
34. M. A. Riehle, Y. Fan, C. Cao, M. R. Brown, *Peptides* **27**, 2547 (2006).
35. R. F. Mercado-Curiel *et al.*, *BMC Microbiol.* **6**, 85 (2006).
36. C. F. Bosio, R. E. Fulton, M. L. Salasek, B. J. Beaty, W. C. Black, *Genetics* **156**, 687 (2000).
37. D. W. Severson, J. K. Meece, D. D. Lovin, G. Saha, I. Morlais, *Insect Mol. Biol.* **11**, 371 (2002).
38. R. C. Jansen, J. P. Nap, *Trends Genet.* **17**, 388 (2001).


39. The *Aedes aegypti* genome sequencing project at the microbial sequencing centers and VectorBase was funded by National Institute of Allergy and Infectious Diseases (NIAID) contracts HHSN266200309D266030071,

HHSN266200400001C, and HHSN266200400039C and was supported in part by NIAID grants UO1 AI50936 (D.W.S.), RO1 AI059492 (A.S.R., G.D.), 5 R01 AI61576-2 (G.D.), and R37 AI024716 (A.S.R.) and by Swiss National Science Foundation grant SNF 3100A0-112588/1 (E.M.Z.). We acknowledge the excellent work of the Broad Genome Sequencing Platform and the Venter Institute Joint Technology Center. We thank C. Town, N. Hall, and E. Kirkness for critical comments and the *Aedes aegypti* research community for their enthusiastic support and willing assistance in this project. On 1 October 2006 The Institute for Genomic Research merged with the J. Craig Venter Institute. The *Ae. aegypti* genome can also be accessed at VectorBase (http://aaegypti.vertorbase.org).

**Supporting Online Material**
www.sciencemag.org/cgi/content/full/1138878/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S23
References

15 December 2006; accepted 7 May 2007
Published online 17 May 2007;
10.1126/science.1138878
Include this information when citing this paper.

REPORTS

# Do Vibrational Excitations of $CHD_3$ Preferentially Promote Reactivity Toward the Chlorine Atom?

**Shannon Yan,[1] Yen-Tien Wu,[1] Bailin Zhang,[1*] Xian-Fang Yue,[1†] Kopin Liu[1,2‡]**

The influence of vibrational excitation on chemical reaction dynamics is well understood in triatomic reactions, but the multiple modes in larger systems complicate efforts toward the validation of a predictive framework. Although recent experiments support selective vibrational enhancements of reactivities, such studies generally do not properly account for the differing amounts of total energy deposited by the excitation of different modes. By precise tuning of translational energies, we measured the relative efficiencies of vibration and translation in promoting the gas-phase reaction of $CHD_3$ with the Cl atom to form HCl and $CD_3$. Unexpectedly, we observed that C–H stretch excitation is no more effective than an equivalent amount of translational energy in raising the overall reaction efficiency; $CD_3$ bend excitation is only slightly more effective. However, vibrational excitation does have a strong impact on product state and angular distributions, with C–H stretch-excited reactants leading to predominantly forward-scattered, vibrationally excited HCl.

Several decades of experimental and theoretical molecular collision studies culminated in the formulation of Polanyi's rules of reaction dynamics (*1*). For reactions of an atom with a diatomic molecule, the rules predict the efficiency of reactant vibrational and translational energy in driving reactions over barriers; namely, vibration can be more effective than translation for a barrier located late along the reaction coordinate, and the reverse is true for reactions with early barriers. An extension of the rules to reactions of polyatomic species becomes ambiguous as a result of the higher degrees of freedom associated with multiple types of vibrational motion. Thus, one may ask: Are different vibrational modes equivalent in their capacity to promote a polyatomic reaction?

In recent years, the issue of mode-specific or bond-selective chemistry (*2*–*5*) has been the subject of several pioneering investigations, for which the reaction of the Cl atom with methane is becoming the benchmark (*6*–*19*). For example, Simpson *et al*. found that one-quantum excitation in the antisymmetric stretch ($\nu_3$) mode of $CH_4$ increases the reaction rate by a factor of ~30 (*10*). On the other hand, Zhou *et al*. observed a mere threefold reactivity enhancement for one-quantum excitation of bending ($\nu_4$) or torsional ($\nu_2$) modes of $CH_4$ and $CD_4$ (*18*), in contrast to 200-fold and 80-fold enhancements measured earlier (*12*, *13*). Further experiments (*17*) and a quasiclassical trajectory calculation (*20*) supported the results of Zhou *et al*. Moreover, Yoon *et al*. found that excitation of the $\nu_1 + \nu_4$ symmetric stretch-bend combination mode of $CH_4$ enhances reactivity toward the Cl atom roughly twice as much as does the nearly isoenergetic excitation of the antisymmetric combination $\nu_3 + \nu_4$, which itself promotes a 10-fold rate enhancement over ground-state methane (*6*). In a similar study, Yoon *et al*. observed a sevenfold reactivity increase of $CH_3D$ when the symmetric, rather than antisymmetric, C–H stretching mode was initially excited (*8*). All these experiments, however, were performed at a fixed translational or collision energy ($E_c$); thus, the enhanced reactivity refers to a comparison with the ground-state reaction at the same $E_c$. As elegant as these experiments are, it remains uncertain whether vibrational motion is more effective in driving this reaction than translation.

We report here a series of experiments aimed to resolve this uncertainty for the $Cl + CHD_3 \rightarrow HCl + CD_3$ reaction. We first studied the ground-state reaction over a wide energy range from the threshold to about 20 kcal/mol of excess energy. Experiments were then performed for the reaction with C–H stretch-excited $CHD_3$, again over a range of initial $E_c$. To refine the comparison, we also present the results for the bend- and/or torsion-excited reactants. We performed all measurements under single-collision conditions, using the rotatable, crossed molecular-beam apparatus described previously (*21*, *22*). The Cl beam was generated by a pulsed high-voltage discharge of ~4% $Cl_2$ seeded in a pulsed supersonic expansion of either Ne or He at 6 atm. The $CHD_3$ beam was also produced by pulsed supersonic expansion of either pure $CHD_3$ or ~20% $CHD_3$ seeded in $H_2$ (for acceleration) at 5 atm. Both beams were collimated by double skimmers and crossed in a differential-pumped scattering chamber. $E_c$ was tuned by varying the intersection angle of the two molecular beams. A pulsed ultraviolet laser that was operated near 333 nm probed the ground-state $CD_3$ product via (2 + 1) resonance-enhanced multiphoton ionization, and a time-sliced velocity imaging technique mapped the recoil vector of the $CD_3^+$ ion (*21*). For studies with C–H stretch-excited reactants, an infrared (IR) laser was used to excite $CHD_3$ directly in front of the first skimmer (*19*). For reactions with bend-excited reactants, a heated pulsed valve for thermal excitation was used instead (*18*).

Figure 1 shows two typical raw images, with and without the IR-pumping laser, of the probed $CD_3(v = 0)$ products at $E_c$ = 8.9 kcal/mol. Superimposed on the images are the scattering directions; the 0° angle refers to the initial $CHD_3$ beam direction in the center-of-mass frame. Thanks to the time-sliced velocity imaging approach, even the raw data can be easily interpreted by inspection. Whereas the IR-off image is dominated by a side-scattered structure, the IR-on image exhibits two distinct ringlike features reflecting the impact of C–H stretch excitation on the reaction dynamics (*23*). A sharp forward peak now appears in the inner ring, and additional broad-scattered products form the outer ring. The

[1]Institute of Atomic and Molecular Sciences, Academia Sinica, Post Office Box 23-166, Taipei, Taiwan 10617. [2]Department of Chemistry, National Taiwan University, Taipei, Taiwan 10617.

*Present address: Department of Chemistry, Wayne State University, Detroit, MI 48202, USA.
†Present address: Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China.
‡To whom correspondence should be addressed. E-mail: kliu@po.iams.sinica.edu.tw

# III.4. Artículo 4

## On the fixation of transposable elements in the genome of the malaria mosquito *Anopheles gambiae.*

(*Manuscript in preparation*)

**Title**

On the Fixation of Transposable Elements in the Genome of the Malaria Mosquito *Anopheles gambiae*

**Authors**

J Tubio[1,3] M Tojo[2], E. Valadé[3], H Naveira[4], J Costas[5] & N Besansky[6].

[1] Servicio de Hematoloxía. Complexo Hospitalario Universitario de Santiago. Santiago de Compostela, 15706. Spain.

[2] Servicio de Anatomía Patolóxica. Complexo Hospitalario Universitario de Santiago. Santiago de Compostela, 15706. Spain.

[3] Departamento de Xenética. CIBUS. Universidade de Santiago de Compostela. Santiago de Compostela, 15706. Spain.

[4] Departamento de Biología Celular y Molecular. Universidade de A Coruña. A Coruña, 15001. Spain.

[5] Fundación Pública Galega de Medicina Xenómica. Complexo Hospitalario Universitario de Santiago. Santiago de Compostela, 15706. Spain.

[6] University of Notre Dame. Notre Dame, IN 46556. USA.

**Abstract**

The transposable element (TE) complement of the genome of *Anopheles gambiae* contains a significant proportion of inactive degenerated elements, whereas the genome of *Drosophila melanogaster* is characterized by a high degree of homogeneity and a lack of divergent elements. Taking into account the selfish essence of transposable elements, natural selection is expected to limit the abundance of TEs in a genome. Each transposable element copy has an indeterminately probability of fixation which depends on selective pressure, population size and recombination frequencies of the surrounding genome. The overall preponderance of fragmentary copies of retrotransposons (including Solo-LTRs) in the genome of *Anopheles gambiae* indicates that most of them have resided for a long time in the genome, and our expectation is that they are probably fixed. In an effort to test this hypothesis, in this work we report the occupancy rate of three antique copies of retrotransposons of the genome of *Anopheles gambiae* (named *locus1*, *locus2* and *locus3*). The presence of these copies was tested in 163 mosquitoes of *Anopheles gambiae*, belonging to molecular forms M and S and recovered from west and east Africa populations, and 20 mosquitoes of *Anopheles arabiensis*, from two East Africa populations.

According to our expectations, we found a high occupancy rate of 1.0 for the three TE loci analyzed in the genome of *Anopheles gambiae*. Two of these loci are also present in *Anopheles arabiensis* and, surprisingly, one of them was detected in all the individuals analyzed, being probably fixed in this sister species. This work represents the first report of transposable elements fixed in *Anopheles gambiae*, and provides the preliminary data for future genome-wide analysis. We also discuss the evolutionary forces probably involved in the fixation of these loci, attending to characteristics of the genome and the evolutionary biology of this species.

## Introduction

*Anopheles gambiae* and *Anopheles arabiensis* are two of the seven morphologically indistinguishable species that comprise the complex of *Anopheles gambiae*, and are considered the two most important vectors for human malaria transmission in Africa. *Anopheles arabiensis* is the most likely ancestral species of the complex (Ayala and Coluzzi, 2005). These two sibling species are widespread and extensively sympatric, and are outcomes of relatively recent speciation processes where paracentric inversions were probably involved. Hybrid females are fertile between all of the species pairs of the complex, therefore there is a potential gene flow that persists up to present. In nature, F1 hybrids between *Anopheles gambiae* and *Anopheles arabiensis* are detected rarely (0.02-0.76%), revealing important premating isolating mechanisms between sympatric populations of these two species (Temu et al., 1996; Toure et al., 1998).

In West Africa, the species *Anopheles gambiae* is composed of sympatric taxa that are at least partially reproductively isolated. These taxa were initially described based on characteristic combinations of paracentric inversions of chromosome 2, between which the expected karyotype frequencies assuming random mating were not found. This led to the splitting of *Anopheles gambiae* into five chromosomal forms (Coluzzi et al., 1985). These chromosomal forms display different ecological tolerances and behaviors. The ecological and chromosomal differences were suggestive of reproductive isolation between forms, but actually they may be explained by differential selection on inversions conferring adaptations to alternative ecological niches exploited by a polymorphic but panmictic *Anopheles gambiae* (Krzywinski & Besansky, 2003). Nevertheless, molecular discontinuities within *Anopheles gambiae* were later discovered in the ribosomal DNA region (Favia et al., 1997) and two molecular forms, named M and S, were defined in the species (della Torre et al., 2001). The tentative splitting of *Anopheles gambiae* into two molecular forms was supported by the virtual absence of hybrid rDNA genotypes in nature. The two forms are characterized by a high level of gene flow restriction, by a largely overlapping geographical and temporal distribution, and by a low degree of genetic differentiation (della Torre, Tu & Petrarca, 2005).

Transposable elements are ubiquitous components of eukaryotic genomes, but differ widely in their abundance (Lynch & Connery, 2003). The vast majority of mobile elements have deleterious effects on the host genome, and usually they are efficiently eliminated by

selection (Charlesworth, Sniegowski & Stephan, 1994). In a previous work Tubío, Naveira & Costas (2005) reported a slower turnover rate of retrotransposons in the genome of *Anopheles gambiae* with respect to *Drosophila melanogaster*, characterized by an overrepresentation of inactive-degenerated divergent elements and few active homologous copies. This overall preponderance of inactive divergent copies in all the families of retrotransposons indicates that most insertions have resided for a long time (as pseudogenes) in the genome of *Anopheles gambiae* and they are now probably fixed (Tubío, Costas & Naveira, 2004).

In this work we report transposable elements fixed in the genome of the species *Anopheles gambiae*, also shared with the species *Anopheles arabiensis*, and we discuss the evolutionary forces that could explain this fact based on (1) low recombination rate in heterochromatic regions (Bartolomé & Maside, 2004; Maside et al., 2005); (2) suppressed recombination in inversions (Sniegowsky and Charlesworth, 1994), and (3) low effective population size (Charlesworth, Sniegowsky & Stephan, 1994).

## Methods

### Study sites and mosquito collections

In Kenya (East Africa), the study sites include Asembo Bay (in the text referred to as Asembo), located on the shores of Lake Victoria in western Kenya, and Jego, located 700km away, on the coast of the Indian Ocean near the Tanzanian border. Mosquitoes of the species *A. gambiae* belonging to molecular form S were collected from both sites in 1987. The DNA samples analyzed in this study include 46 mosquitoes from Asembo and 19 mosquitoes from Jego. In addition, DNA samples from 10 mosquitoes of *A. arabiensis* collected in Asembo were also analyzed. Both populations, Asembo and Jego, are separated by the Great Rift Valley, which represents a restrictive barrier to gene flow within the species *Anopheles gambiae* (Lehmann et al. 2003). In Burkina Faso (West Africa), the study site was Goundry, where mosquitoes were collected in 2001. The DNA samples analyzed include 50 mosquitoes of the species *A. gambiae* form M, 48 of *A. gambiae* form S and 10 mosquitoes of *A. arabiensis*.

### *Anopheles gambiae* complex ID-PCR

The *Anopheles gambiae* complex comprises seven cryptic species. We performed the ID-PCR developed by Scott et al (1993) to distinguish individuals of *A. gambiae* and *A. arabiensis*. This PCR also allowed us to detect any possible trace of DNA contaminated between species. Briefly, 1.0 μM primer forward *UN* (GTG TGC CCC TTC CTC GAT GT), 1.0 μM primer reverse *GA* (CTG GTT TGG TCG GCA CGT TT), 1.0 μM primer reverse *AR* (AAG TGT CCT TCT CCA TCC TA), 1.0mM $Mg^{+2}$, 0.2mM dNTPs, 1x Taq buffer and 1U Taq polymerase. The PCR cycle conditions were as follows: 1 cycle 95ºC/10min; 30 cycles 95ºC/30sec, 50ºC/30sec, 72ºC/30sec; 1 cycle 72ºC/5min. Primers created fragments of 390bp for *A. gambiae* and 315bp for *A. arabiensis*.

### *Anopheles gambiae* ribosomal DNA type assessment

The distinction between the two molecular forms of *A. gambiae* was made following the method of Fanello et al. (2002). This method use the restriction enzyme *HhaI* to digest the PCR product obtained by the Scott et al. assay (commented previously). Briefly, 0.5μl of

*HhaI* restriction enzyme was added to 10µl of PCR product, allowing an incubation of 24 hours at 37ºC. After loading in agarose ethidium bromide gel, molecular form M revealed a 397bp band and molecular form S revealed 225bp and 110bp bands.

**Selection of TE loci**

We selected the TE loci under two premises: (1) they must be antique, and (2) they must display short size. From a database of 642 TEs previously identified by our group (Tubío, Naveira and Costas, 2005), we filter all the sequences according to these premises and selected and initial set of 10 TE loci with best score. After several PCR tests, we finally chose those three TEs with better amplification results. These TE loci are detailed in table 1.

**Primer design and Genotyping**

Two different PCRs and, therefore, two different sets of primers were designed per TE locus. The first set of primers, named F1 and R1 (forward1 and reverse1, respectively), were designed outside the TE copy and flanking it. For the second set of primers, named F1 and R2 (forward1 and reverse2, respectively), the reverse primer was designed inside the copy. All primers were designed using the program GeneFisher (Giegerich et al., 1996), using default parameters. PCR reactions were as follows: 0.6 µM primer forward, 0.6 µM primer reverse, 1.5 mM $Mg^{+2}$, 0.16 mM dNTPs, 1x Taq buffer and 1.25 U Taq polymerase were mixed in a total reaction volume of 25µl. The PCR cycle conditions were as follows: 1 cycle 95ºC/5min; 35 cycles 95ºC/30sec, annealing temperature/30sec, 72ºC/45sec; 1 cycle 72ºC/10min. Genotyping was carried out in a 3730 Applied Biosytems sequencer, using GeneScan 500 LIZ ladder (Applied Biosystems). The sequences of the primers, annealing temperatures and expected amplicons sizes are detailed in table 2.

**Sequencing**

Briefly, the amplicons were extracted from a 3% agarose gel, and purified using QIAquick Gel Extraction Kit (Qiagen). Purified amplicons were concentrated from the eluted volume of 80µl to a final volume of 8µl, of wich 4µl were used for each sequencing PCR (forward

and reverse PCR reactions) using Bigdye v3.0 (Applied Biosystems). Dyes were removed using an ethanol-based protocol, and finally PCR products were diluted in deionized formamide. Sequencing was carried out in a 3730 Applied Biosytems sequencer.

## Results

### High occupancy rate of transposable elements in the genome of *A. gambiae*

For the three TE loci under study we found high occupancy rates of 1.0 (table 3). We confirmed the presence of the three loci by genotyping after two different sets of polymerase reactions per locus. The expected amplicon size for the presence of each locus, reported in table 3, was obtained in all the 163 cases for the three loci analyzed.

### Fixation of TEs in the genome of *Anopheles gambiae*

These three TE loci are present in all the populations and subpopulations analyzed of *A. gambiae* by genotyping. Furthermore, we verified this observation by sequencing each of these three TE loci in all the populations and subpopulations, confirming the presence of the TE loci in all of them. The sequence of the junctions were >99% identical to the reference sequence genome of the PEST strain (Holt et al., 2002). In addition, the comparisons of the entire nucleotide sequences of these TE loci with the sequences available in the sequenced genome of the PEST strain, are consistent with the fixation of these elements in *Anopheles gambiae*.

Figures 1, 2 and 3 show the alignment of *locus1*, *locus2* and *locus3*, respectively, among the populations studied.

### Shared TE loci between *Anopheles gambiae* and *Anopheles arabiensis*

We obtained clear evidence that *locus1* and *locus3* are present in *Anopheles arabiensis*. All *Anopheles arabiensis* specimens analyzed present the expected amplicon size for TE *locus1*, revealing an occupancy rate of 1.0 (20/20). Nevertheless, for *locus3* amplification was only successful when using primers F1 and R2, and only few samples of *Anopheles arabiensis* showed robust amplification by PCR using the same PCR parameters as for *Anopheles gambiae*, precluding estimates of the occupancy rate for *locus3* in this sister species without re-optimization of the PCR conditions is necessary. Nevertheless, robust amplification of *locus3* in some samples (using primers F1 and R2) allowed us to obtain the sequence of the junctions and, therefore, to certify the presence of this TE loci in *Anopheles arabiensis*.

**Location of the fixed loci within the genome of *Anopheles gambiae***

*Locus1* and *locus3* are most probably located in the euchomatine: ENSEMBL shows nearby genes for *locus1* in the chromosome 3R (ENSANGT00000015724 and ENSAGT00000015733) and for *locus3* in the chromosome 2L (ENSANGT0000004349 and ENSAGT00000015880). Nevertheless, the *locus2* was not mapped to a chromosome.

## Discussion

The long term persistence of transposable elements in eukaryotic genomes has been explained by models of selfish DNA evolution, which postulate a strong competition among functional copies of a given family within the host genome (Hickey, 1982), although more subtle interactions between different TE families may play a role as well (Leonardo and Nuzhdin, 2002). Natural selection is expected to put limits to the total copy number of transposable elements because of deleterious gene mutations and chromosomal rearrangements promoted by them. Thus, each specific retrotransposon insertion has an indeterminately low probability of fixation, which depends on selective pressure, population size and recombination frequencies at the surrounding genome (Charlesworth et al., 1997).

Lerat, Rizzon and Biémont (2003) reported that the absence of divergence among copies of LTR retrotransposons observed in the genome of *Drosophila melanogaster* could result from a rapid turnover that eliminates transposable element copies as soon as they become inactive. Results obtained previously by Kaminker et al. (2002), based on average pairwise comparisons among insertions of the same family of retrotransposons, are in agreement with this observation. The study of recent insertion profiles for retrotransposons in the genome of *Drosophila melanogaster* (Bergman & Besasson, 2007) is in agreement with a rapid turn over rate. In contrast, Tubío, Naveira & Costas (2005) have reported that the genome of the mosquito *Anopheles gambiae* displays an overall preponderance of fragmentary (including solitary LTRs) indicating that most insertions have resided for a long time in the genome and now they are probably fixed (Tubío, Naveira & Costas, 2004).

### Fixation of transposable elements in the genome of *Anopheles gambiae*

In this paper, we provide clear evidence that the three loci analyzed are fixed in *Anopheles gambiae*. The overall populations and subpopulations studied are a reliable representation of the entire population structure of the species, considering mosquitoes from M and S molecular forms in West Africa, the barrier to gene flow represented by the Great Rift Valley in East Africa, and the possible role of "isolation by distance" between populations separated by >6000 Kilometers (Lehmann et al., 2003). The evidence of fixation is revealed by the fact that all the individuals analyzed of the species *Anopheles gambiae* present the expected amplicon size. Furthermore, sequencing of at least one amplicon of

each population is in agreement with fixation. In addition, it was possible to detect and sequence two of the loci analyzed in *Anopheles arabiensis*, and there is evidence of a possible fixation of *locus1* in this sister species, being present in 20/20 mosquitoes studied from West and East Africa populations.

Fixation of transposable elements is the genome of *Anopheles gambiae* is congruent with previous expectations (Tubío, Costas & Naveira, 2004), by the observation of significant differences between the overall patterns of turnover rate displayed by *Anopheles gambiae* with respect to *Drosophila melanogaster*. By the line of models of selfish DNA evolution, we expect that the persistence of inactive-antique copies in the genome of *Anopheles gambiae* could be due to a reduction of selective pressure, mainly by insertion in lower recombination genomic regions, and/or due to a reduced effective population size, that could characterized this species.

**Insertion within heterochromatin**

Transposable element loci are more likely to reach higher frequencies in those regions of the genome characterized by a lower recombination rate (Bartolomé et al., 2002). In *D. melanogaster* the insertion of transposable elements in heterochromatic regions let to the fixation of some transposable elements. Bartolome & Maside (2004) found by PCR that transposable elements of *D. melanogaster* tend to fixation in non-recombining regions, particularly in chromosome 4 (where recombination levels are very low). These results were later supported by in-situ hybridization, when the fixation of two transposable element loci in the beta-heterochromatin of *D. melanogaster* was detected (Maside, Assimacopoulos & Charlessworth, 2005).

Holt et al. (2002) reported that the transposable elements of the genome of *A. gambia*e constitute about 60% of the heterochromatic component, being highly fragmented. Within the euchromatic component of the genome, repeat density is higher near the centromeres, lowest in the middle of chromosome arms, and somewhat elevated near the telomeres. Holt et al. (2002) pointed out that transposable element distribution is consistent with the hypothesis that densities are higher in parts of the genome where recombination rates are lowest. By the line of this suggestion, we searched the location of each TE-loci analyzed in ENSEMBL. The results let us to hypothesize that the 3R and 2L loci are euchromatic based on nearby genes and the fact that the positions do not abut the centromere. We were

not be able to determine wether the ubication of the *locus2* is euchromatic or heterochromatic, as it was not even mapped to a chromosome, but this, by itself, suggests that it lies on a short scaffold filled with repetitive DNA.

**The effect of hitchhiking associated to adaptative chromosomal inversions**

Chromosomal inversions are thought to play important roles in ecological differentiation in speciation, by suppressing recombination between hetero-karyotypes and by stabilizing adaptative combinations of genes (Powell et al., 1999; Ayala & Coluzzi, 2005). Within the subfamily *Anophelinae* gene order has been extensively shuffled by chromosomal inversions (Sharakhov et al., 2002). The basal pool of polymorphic chromosomal arrangements provides the genetic variability necessary to exploit efficiently niches with different ecological conditions, before the fixation of those more efficient configurations. Along the way of this evolutionary process several loci associated with the selected inversion could fix in a population due to a hitch-hiking effect.

A well-known example of the importance of these arrangements in the evolution of species comes from mosquitoes the 2La paracentric inversion of the *Anopheles gambiae* complex (Ayala & Coluzzi, 2005). *Anopheles arabiensis* is considered the most likely ancestral species of the complex of *A. gambiae*. It exhibits a fixed second-chromosome arrangement (2La), which is though to be ancestral because it is also present in other species groups of mosquitoes such as the *Anopheles subpictus* complex, where it is fixed in at least one of the siblings. *Anopheles gambiae sensu stricto* is the only member of the complex in which the 2La inversion is polymorphic. The adaptative value of the 2La inversion in the *Anopheles gambiae* complex seems to be unquestionable. The frequently of this inversion is nonrandom with respect of aridity (Powell et al., 1999), being absent in humid forest and fixed in arid Sahel. It has been reported that low recombination areas within a genome are potential hotspots for the accumulation of transposable elements. The absent or reduced recombination in these genome areas makes possible a relaxation natural selection against transposable elements, because the potential deleterious effects of ectopic recombination are highly reduced. Stump et al. (2007) have recently reported that the recombination rates for hetero-karyotypes (2La/+) in the malaria vector *Anopheles gambiae* is <0.5 cM/Mb, whereas the recombination rates in homo-karyotypes was estimated in a value of ~2.0 cM/Mb.

Results consistent with the hypothesis of hitchhiking associated to inversions were reported in the genome of *Drosophila melanogaster*, where Sniegowski and Charlesworth (1994) found that transposable elements occupy chromosomal sites at significantly higher frequencies within the inversions In(3R)Mo and In(3R)K. For these two inversions, higher copy numbers of transposable elements were accompanied by significant increases in element frequencies at occupied sites. The authors pointed out that hitchhiking of transposable element insertions might account for the observed increases in TE copy number. More recently, Depaulis et al. (2000) have also found evidence for hitchhiking effects with recombination of loci distant from proximal breakpoint of the inversion In(2L)t of *Drosophila*. It is clear, therefore, that chromosomal inversions largely inhibit recombination and may be associated with the effect of positive selection on linked loci.

A suppressed-recombination model of chromosome speciation was proposed by Coluzzi (reviewed by Ayala & Coluzzi, 2005) to account for the speciation patterns showed by the *Anopheles gambiae* complex. Under hitchhiking, the spread of a TE loci located within an adaptative inversion will presumably increase, possibly to fixation, if the effect of the TE loci is not deleterious. Ayala and Coluzzi (2005) pointed out that among the nearly 500 known members of the genus *Anopheles*, there are no fewer than 170 cryptic taxa belonging to 30 complexes of closely related species, and most siblings are outcomes of recent speciation processes detected by paracentric inversions. Therefore, the crossover suppression provided by the chromosome rearrangements could be considered as an alternative explanation for the probably overall higher occupancy rates of inactive-divergent retrotransposons in the genome of *A. gambiae*. This explanation supposes that those loci were associated to ancient adaptive inversions (probably different to the paracentric inversion pattern displayed today) and, as consequence of the suppressed deleterious ectopic exchange, they became fixed linked to those ancient inversions.

Taking into account that the three fixed are antique (probably most than the beginning of the speciation processes of the complex that became ~4000 years ago) it is not surprising that *Anopheles gambiae* and *arabiensis* share at least two of the three TE loci analyzed. Furthermore, the most likely explanation for the virtual absence of *locus2* in *A. arabiensis* could be primer mismatching due to nucleotide differences between both genomes. Although, other possibility is that the magnitude of the forces tending to remove elements from ancestral inversions is enough to change the pattern of TE loci left by common

descent. Nevertheless, it is difficult to resolve if the hypothesis of hitchhiking is correct due to the low information available about the ancestral rearrangements in the genus Anopheles. Xia, Sharakhova and Sharakhov (2007), based on physical maps of the outgroup species Anopheles funestus and Anopheles stephensi and using bioinformatics programs, have reconstructed the inversion history of the complex estimating the chromosomal arrangements which are more likely to be ancestral in the complex of Anopheles gambiae, but the information reported is enough to understand the genome configuration of an ancient species of the Complex.

**Population structure**

Patterns of variability within species can also be shaped by demographic history, that is, the structure and size of populations over time. The effects of demography and selection/hitchhiking can be distinguished by considering data from multiple loci (Andolfatto, 2001). Demography has similar effect on the whole genome, whereas the effects of selection and hitchhiking are locus-specific or restricted to the region within arrangements.

On a continental scale *Anopheles gambiae* does not appear to exist as a single, genetically undifferentiated population, but it can be roughly subdivided in at least to major population groups, separated by the Rift Valley. In addition, in West Africa distinct subpopulations exists at local scale, named molecular forms M and S, with restriction to gene flow. Current demographic data for *Anopheles gambiae* ruled out bottlenecks along the recent evolutionary history of the species, except for the East of Kenya (Lehmann et al., 2000). In addition, according to Lehman et al. (1998) an efficient long-term population size is probably measured in hundreds of thousands and hence does not support recent expansions of this species from small populations. Ayala & Coluzzi (2005) proposed that *Anopheles arabiensis*, the presumable most ancestral species of the complex of *Anopheles gambiae*, descended from a *Pyretophorus* species from the Arabian Peninsula that was zoophilic and exophilic. Unfortunately, the absence of demographic information for this ancestral-like species, and the low number of TE loci analyzed, limits an interpretation for the effect of demographic history on a expected high occupancy rate of retrotransposons.

**Bibliography**

Andolfatto, P. (2001). "Adaptive hitchhiking effects on genome variability." Curr Opin Genet Dev 11(6): 635-41.

Ayala, F. J. and M. Coluzzi. 2005. Chromosome speciation: humans, Drosophila, and mosquitoes. Proc Natl Acad Sci U S A 102 Suppl 1: 6535-42.

Bartolome, C., X. Maside, et al. (2002). "On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster." Mol Biol Evol 19(6): 926-37.

Bartolome, C. and X. Maside (2004). "The lack of recombination drives the fixation of transposable elements on the fourth chromosome of Drosophila melanogaster." Genet Res 83(2): 91-100.

Bergman, C. M. and D. Bensasson (2007). "Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in Drosophila melanogaster." Proc Natl Acad Sci U S A 104(27): 11340-5.

Coluzzi, M., V. Petrarca, et al. (1985). "Chromosomal inversion intergradation and incipient speciation in *Anopheles gambiae*." Boll. Zool. 52: 45-63.

Charlesworth, B., P. Sniegowski, et al. (1994). "The evolutionary dynamics of repetitive DNA in eukaryotes." Nature 371(6494): 215-20.

Charlesworth, B., C. H. Langley, et al. (1997). "Transposable element distributions in Drosophila." Genetics 147(4): 1993-5.

della Torre, A., C. Fanello, et al. (2001). "Molecular evidence of incipient speciation within Anopheles gambiae s.s. in West Africa." Insect Mol Biol 10(1): 9-18.

della Torre, A., Z. Tu, et al. (2005). "On the distribution and genetic differentiation of Anopheles gambiae s.s. molecular forms." Insect Biochem Mol Biol 35(7): 755-69.

Depaulis, F., L. Brazier, et al. (2000). "Selective sweep near the In(2L)t inversion breakpoint in an African population of Drosophila melanogaster." Genet Res 76(2): 149-58.

Fanello, C., F. Santolamazza, et al. (2002). "Simultaneous identification of species and molecular forms of the Anopheles gambiae complex by PCR-RFLP." Med Vet Entomol 16(4): 461-4.

Favia, G., A. della Torre, et al. (1997). "Molecular identification of sympatric chromosomal forms of Anopheles gambiae and further evidence of their reproductive isolation." Insect Mol Biol 6(4): 377-83.

Giegerich, R., F. Meyer, et al. (1996). "GeneFisher--software support for the detection of postulated genes." Proc Int Conf Intell Syst Mol Biol 4: 68-77.

Hickey, D. A. (1982). "Selfish DNA: a sexually-transmitted nuclear parasite." Genetics 101(3-4): 519-31.

Holt, R. A., G. M. Subramanian, et al. (2002). "The genome sequence of the malaria mosquito Anopheles gambiae." Science 298(5591): 129-49.

Kaminker, J. S., C. M. Bergman, et al. (2002). "The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective." Genome Biol 3(12): RESEARCH0084.

Krzywinski, J. and N. J. Besansky (2003). "Molecular systematics of Anopheles: from subgenera to subpopulations." Annu Rev Entomol 48: 111-39.

Lehmann, T., W. A. Hawley, et al. (1998). "The effective population size of Anopheles gambiae in Kenya: implications for population structure." Mol Biol Evol 15(3): 264-76.

Lehmann, T., C. R. Blackston, et al. (2000). "The Rift Valley complex as a barrier to gene flow for Anopheles gambiae in Kenya: the mtDNA perspective." J Hered 91(2): 165-8.

Lehmann, T., M. Licht, et al. (2003). "Population Structure of Anopheles gambiae in Africa." J Hered 94(2): 133-47.

Leonardo, T. E. and S. V. Nuzhdin (2002). "Intracellular battlegrounds: conflict and cooperation between transposable elements." Genet Res 80(3): 155-61.

Lerat, E., C. Rizzon, et al. (2003). "Sequence divergence within transposable element families in the Drosophila melanogaster genome." Genome Res 13(8): 1889-96.

Lynch, M. and J. S. Conery (2003). "The origins of genome complexity." Science 302(5649): 1401-4.

Maside, X., S. Assimacopoulos, et al. (2005). "Fixation of transposable elements in the Drosophila melanogaster genome." Genet Res 85(3): 195-203.

Nene, V., J. R. Wortman, et al. (2007). "Genome sequence of Aedes aegypti, a major arbovirus vector." Science 316(5832): 1718-23.

Powell, J. R., V. Petrarca, et al. (1999). "Population structure, speciation, and introgression in the Anopheles gambiae complex." Parassitologia 41(1-3): 101-13.

Scott, J. A., W. G. Brogdon, et al. (1993). "Identification of single specimens of the Anopheles gambiae complex by the polymerase chain reaction." Am J Trop Med Hyg 49(4): 520-9.

Sharakhov, I. V., A. C. Serazin, et al. (2002). "Inversions and gene order shuffling in Anopheles gambiae and A. funestus." Science 298(5591): 182-5.

Sniegowski, P. D. and B. Charlesworth (1994). "Transposable element numbers in cosmopolitan inversions from a natural population of Drosophila melanogaster." Genetics 137(3): 815-27.

Stump, A. D., M. Pombi, et al. (2007). "Genetic exchange in 2La inversion heterokaryotypes of Anopheles gambiae." Insect Mol Biol 16(6): 703-9.

Temu, E. A., R. H. Hunt, et al. (1997). "Detection of hybrids in natural populations of the Anopheles gambiae complex by the rDNA-based, PCR method." Ann Trop Med Parasitol 91(8): 963-5.

Toure, Y. T., V. Petrarca, et al. (1998). "The distribution and inversion polymorphism of chromosomally recognized taxa of the Anopheles gambiae complex in Mali, West Africa." Parassitologia 40(4): 477-511.

Tubio, J. M., J. C. Costas, et al. (2004). "Evolution of the mdg1 lineage of the Ty3/gypsy group of LTR retrotransposons in Anopheles gambiae." Gene 330: 123-31.

Tubio, J. M., H. Naveira, et al. (2005). "Structural and evolutionary analyses of the Ty3/gypsy group of LTR retrotransposons in the genome of Anopheles gambiae." Mol Biol Evol 22(1): 29-39.

Xia, A., M. V. Sharakhova, et al. (2007). "Reconstructing an inversion history in the Anopheles gambiae complex." Journal of Computational Biology 15(8): 965-980.

**Tables and Figures**

Table 1. TE loci analyzed in the study

| Loci | Family[a] | Scaffold[b] | Chr[c] | Position[d] | div[e] |
|---|---|---|---|---|---|
| *Locus 1* | GYPSY31_AG | AAAB01008980 | 3R | 7409466-7409287 | 0.6 |
| *Locus 2* | 8967 | AAAB01008881 | NP | 208366-208554 | 0.6 |
| *Locus 3* | GYPSY57_AG | AAAB01008960 | 2L | 5250968-5251267 | 0.6 |

[a] Family names according to Tubío, Naveira and Costas (2005).
[b] Scaffold in the genome sequence of PEST strain of Anopheles gambiae.
[c] Chromosomal location.
[d] Position within the scaffold.
[e] Pairwise divergence with respect to the consensus sequence for the corresponding family.

Table 2. Primer sequences and PCR conditions

| *Loci* | | Primers | Size[a] | Temp[b] |
|---|---|---|---|---|
| *Locus 1* | F1 | $^{FAM}$TGG AAT CAC ACT CCA CGA | 491 | 60ºC |
| | R1 | CGG AAG TTG TGC TAG CAA | | |
| | F1 | $^{FAM}$TGG AAT CAC ACT CCA CGA | 275 | 60ºC |
| | R2 | CGT TCT GTT CGC TGT CA | | |
| *Locus 2* | F1 | $^{FAM}$CGT TTT GCC CAT GCC TA | 577 | 56.9ºC |
| | R1 | TTG ACC ATG CGG CAG A | | |
| | F1 | $^{FAM}$CGT TTT GCC CAT GCC TA | 246 | 60ºC |
| | R2 | CGG AAC GGA AAA GAC GA | | |
| *Locus 3* | F1 | $^{FAM}$TCA CAA CAG CCG AAC GA | 600 | 60ºC |
| | R1 | ATG GAC TGC CGC CCT A | | |
| | F1 | $^{FAM}$TCA CAA CAG CCG AAC GA | 185 | 59.1ºC |
| | R2 | TCC CGA TCG GCA CTC A | | |

[a] Size of the expected amplicon.
[b] Annealing temperature.

Table 3. Occupancy rate of the TE loci studied in *Anopheles gambiae*

| | WEST AFRICA | | EAST AFRICA (S form) | |
|---|---|---|---|---|
| *Loci* | M form | S form | Asembo | Jego |
| *Locus 1* | 1.0 (50/50) | 1.0 (48/48) | 1.0 (46/46) | 1.0 (19/19) |
| *Locus 2* | 1.0 (50/50) | 1.0 (48/48) | 1.0 (46/46) | 1.0 (19/19) |
| *Locus 3* | 1.0 (50/50) | 1.0 (48/48) | 1.0 (46/46) | 1.0 (19/19) |

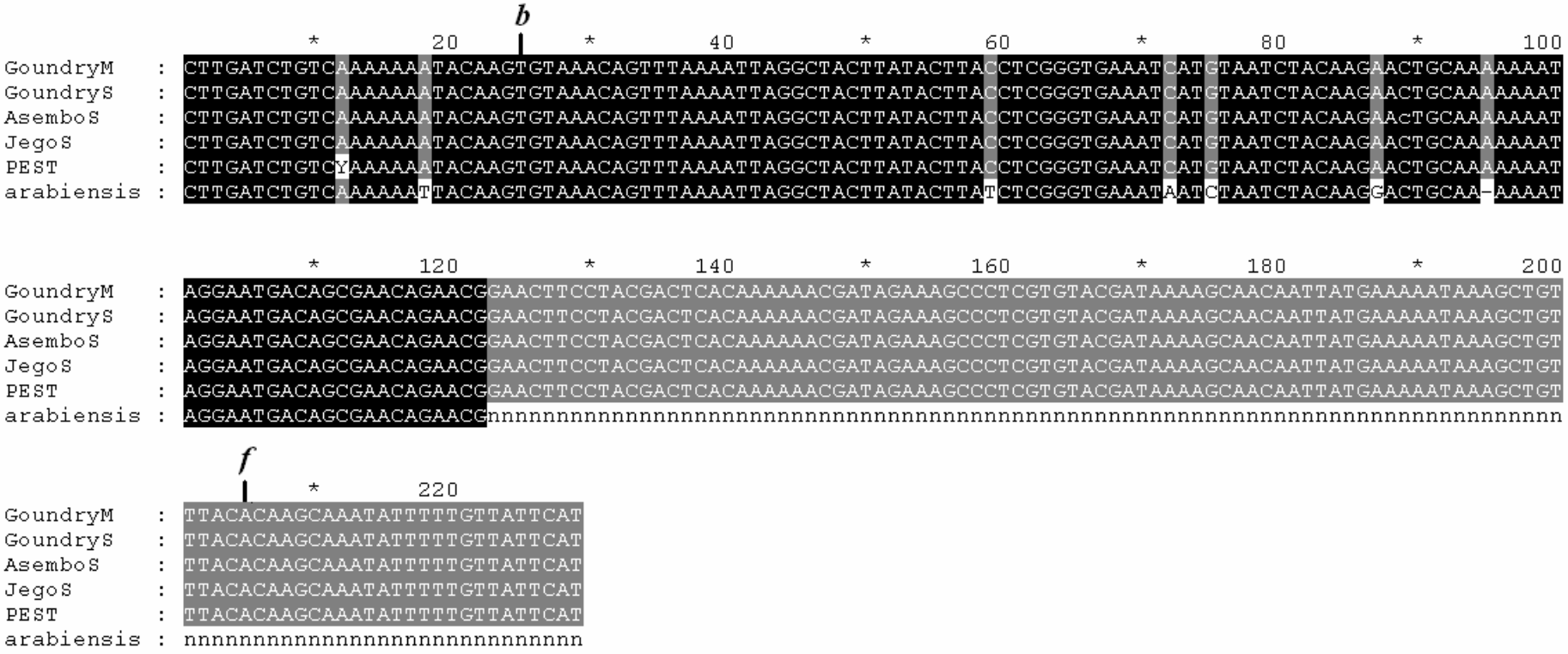



Figure 1. Alignment of nucleotide sequences for *locus1*. Sequence of *locus1* extends from position 25 to position 205; target site duplication (CAAG) extends from positions 21-24 and 206-209. Sequence of the junctions upstream and downstream *locus1* are also displayed (positions 1-20 and 210-229, respectively).

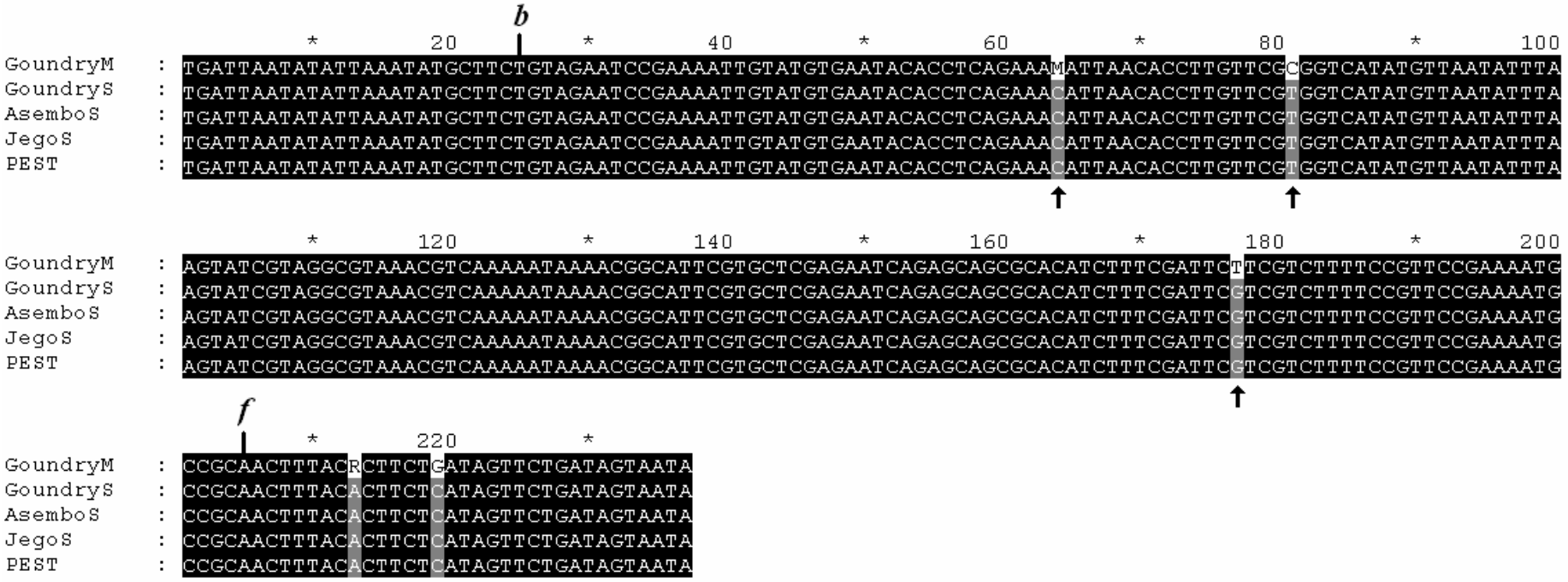



Figure 2. Alignment of nucleotide sequences for *locus2*. Sequence of *locus2* extends from position 25 to position 205; target site duplication (CTTC) extends from positions 21-24 and 206-209. Sequence of the junctions upstream and downstream *locus2* are also displayed (positions 1-20 and 210-229, respectively).

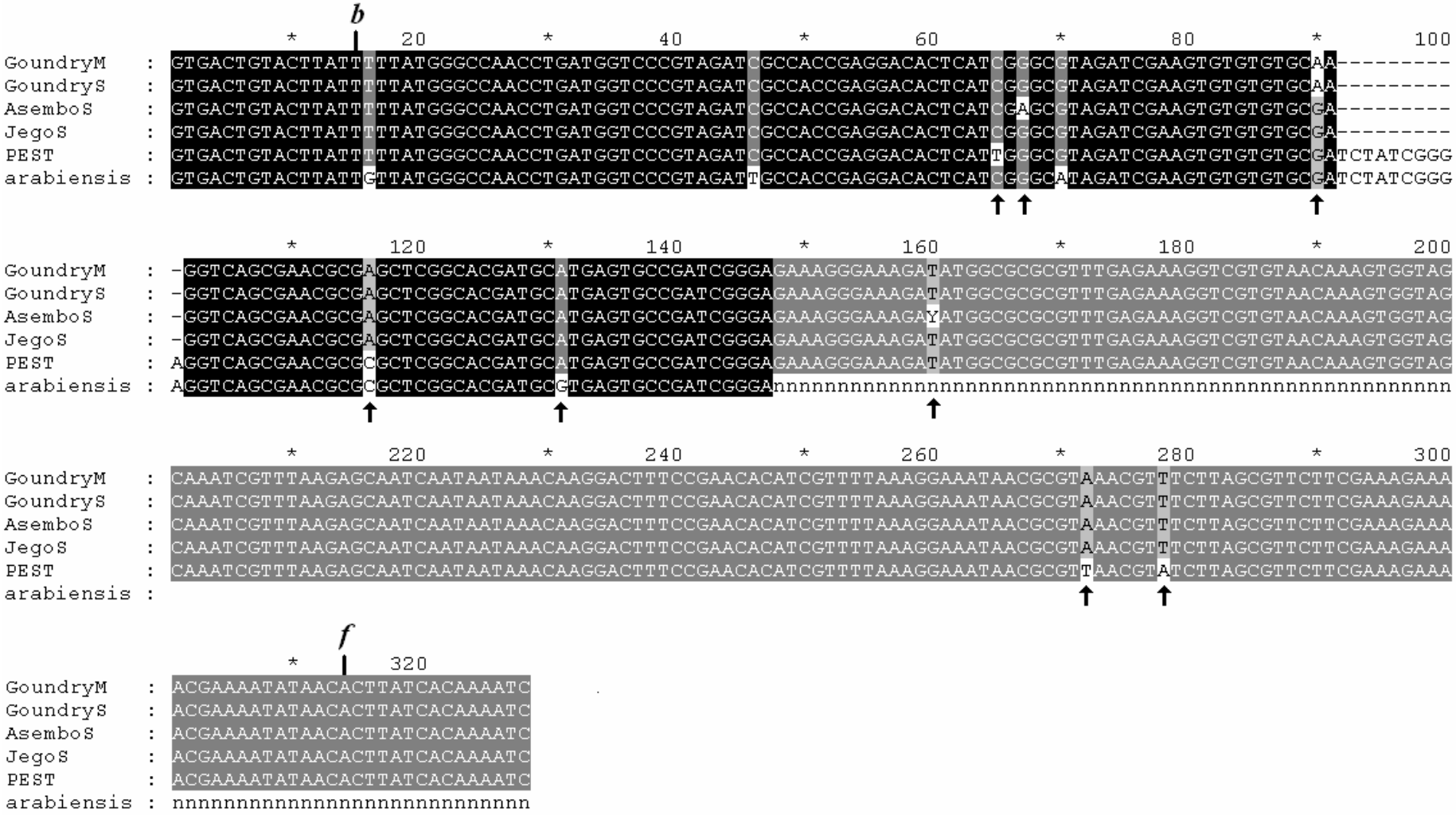



Figure 3. Alignment of nucleotide sequences for *locus3*. Sequence of *locus3* extends from position 15 to position 314; target site duplication (CTTAT) extends from positions 9-14 and 315-319. Sequence of the junctions upstream and downstream *locus3* are also displayed (positions 1-8 and 320-318, respectively).

# IV. Resultados y Discusión

## IV.1. Evolución de los TEs del grupo *Ty3/gypsy* en *Anopheles gambiae.*

Uno de los objetivos generales de esta tesis era caracterizar molecularmente los TEs del grupo *Ty3/gypsy* presentes en el genoma secuenciado de la cepa PEST de *Anopheles gambiae*, como paso previo para conocer la dinámica evolutiva de estos elementos en dicho genoma y para realizar un análisis comparativo con la dinámica que siguen los TEs en el genoma de *Drosophila melanogaster*, entre otras especies. Al iniciar este trabajo, la información disponible en este sentido referida a los TEs del género *Anopheles* era escasa y, a la vez, muy genérica: a penas se habían caracterizado parcialmente las secuencias de un puñado de TEs en el género *Anopheles* (Ke et al., 1996; Biessmann & Mason, 1997; Romans, Bhattacharyya & Colavita, 1998; Biessmann et al., 1999; Grossman et al., 1999; Cook et al., 2000; Hill et al., 2001; Rohr et al., 2002) y, además, la información proporcionada con la secuenciación del genoma de la cepa PEST de *A. gambiae* se ceñía a una aproximación, poco detallada, sobre el porcentaje del genoma del mosquito que ocuparían determinados grupos de TEs (Holt et al., 2002). Concretamente, en lo referido a los TEs de Clase I identificados hasta entonces en el género *Anopheles*, la representación reportada en detalle era muy pobre: tres elementos representantes del *tipo Ty3/gypsy* [*Aste5* y *Afun1* (Cook et al., 2000); *Ozymandias* (Hill et al., 2001)], cinco elementos del *tipo Ty1/copia* [*Amer3*, *Amer6*, *Amer7* y *Aste7* (Cook et al., 2000); *mtanga* (Rohr et al., 2002)], y cinco elementos representantes del *tipo Pao* [*moose* (Biessmann et al., 1999); *Agam10*, *Amer1*, *Aara5* y *Aste12* (Cook et al., 2000)]. De todos estos TEs de Clase I, los únicos retrotransposones identificados en la especie *A. gambiae* eran *Ozymandias*, *mtanga* y *moose*. Así, como primer paso en la investigación se inició la búsqueda de retrotransposones en el genoma de *A. gambiae*.

### IV.1.1. Búsqueda e identificación de familias del grupo *Ty3/gypsy* en el genoma de *Anopheles gambiae.*

Actualmente existen diversos programas informáticos especializados en la búsqueda automática *in silico* de TEs en los genomas (Jordan & Bowen, 2004). Sin embargo, en el momento de iniciar esta investigación los programas disponibles eran pocos y presentaban importantes limitaciones. Así, en este trabajo de investigación, la estrategia general seguida para la localización de las familias de elementos del grupo *Ty3/gypsy* en el

genoma de la cepa PEST de *A. gambiae* ha sido la búsqueda de homología compartida, empleando el programa tBLASTn y tomando como *query* las secuencias *pol*-like de los elementos representativos de cada uno de los seis linajes conocidos del grupo *Ty3/gypsy* presentes en insectos. Este proceso de búsqueda reveló la existencia de, al menos, 73 elementos representativos pertenecientes a diferentes familias presentes en el genoma de la cepa PEST de *A. gambiae* (Tubio, Costas & Naveira, 2004; Tubio, Naveira & Costas, 2005).

Para asignar definitivamente cada una de esas 73 familias a uno de los seis linajes conocidos del grupo *Ty3/gypsy*, se llevó a cabo un análisis filogenético basado en el alineamiento de los dominios RT-RH-INT de cada uno de los elementos representativos de dichas 73 familias y cada uno de los elementos representativos de cada linaje del grupo *Ty3/gypsy* (figura 14). El elevado valor *bootstrap* obtenido (superior al 95% en todos los clados) permitió asignar inequívocamente cada uno de los nuevos elementos encontrados a un linaje del grupo *Ty3/gypsy*.

### IV.1.2. Diversidad y abundancia de retrotransposones en el genoma de *A. gambiae*.

El estudio sobre la representación de los retrotransposones del grupo *Ty3/gypsy* en el genoma de *A. gambiae* reveló una gran diversidad, reflejada en la existencia de una amplia representación de cada uno de los linajes del grupo *Ty3/gypsy* de insectos, con la excepción del linaje de *Osvaldo*. Además, el estudio individual de cada una de las familias de los retrotransposones identificados, nos permitió observar dos interesantes aspectos en su dinámica: por un lado, la abundancia media de copias por familia resultó elevada con respecto a la de *D. melanogaster*; por otro lado, existe una importante diversidad también a nivel de Familia, pues la mayoría de las familias presentan tanto elementos representativos recientes como antiguos, tanto en su forma proviral (completa o incompleta) como en el de LTR solitaria. Este aspecto, tratado con detalle más adelante en el texto, revela un patrón especial en la dinámica de los retrotransposones en el genoma de *A. gambiae*, que lo diferencia claramente de la amplia homogeneidad característica del genoma de *D. melanogaster* (Kaminker et al., 2002; Lerat, Rizzon & Biemont, 2003).

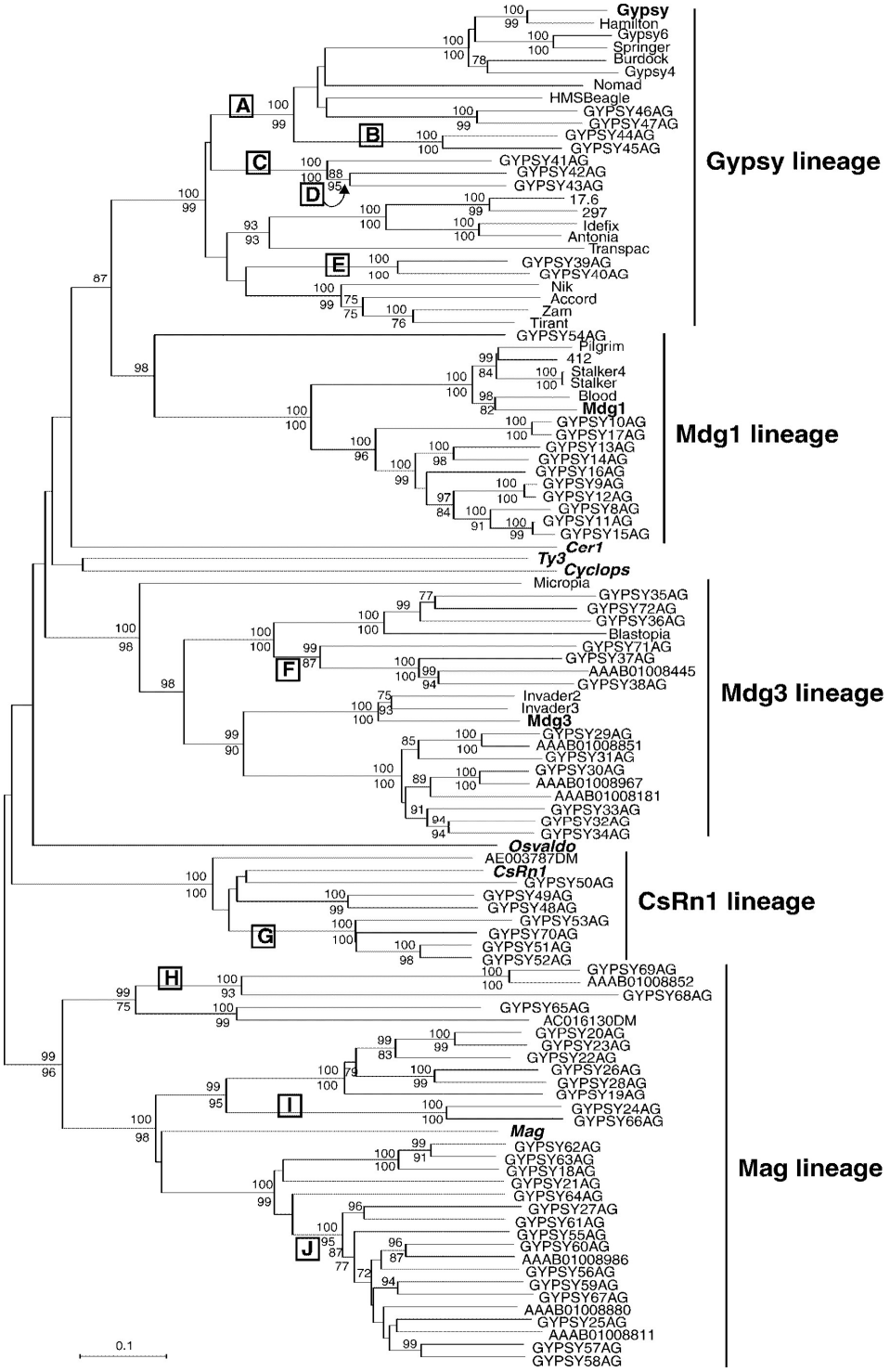



**Figura 14.** Relaciones filogenéticas entre los retrotransposones del grupo *Ty3/gypsy* de *D. melanogaster* y *A. gambiae*, obtenida por NJ tras el alineamiento de los dominios RT-RH-INT. También se muestran los elementos representativos de aquellos linajes no presentes en *D. melanogaster* (cursiva). En las ramas se indican los valores *bootstrap* ≥75% (NJ encima y MP debajo de cada rama). Los nombres de los elementos de *A. gambiae* se indican con el formato propio de Repbase GYPSY##AG (donde # representa un valor numérico), salvo aquellos TEs de *A. gambiae* para los cuales no fue posible obtener una secuencia consenso, que son representados con el nombre del *scaffold* donde fue localizada la copia canónica de la familia (AAAB0100####). Tomada de Tubío, Naveira & Costas (2005).

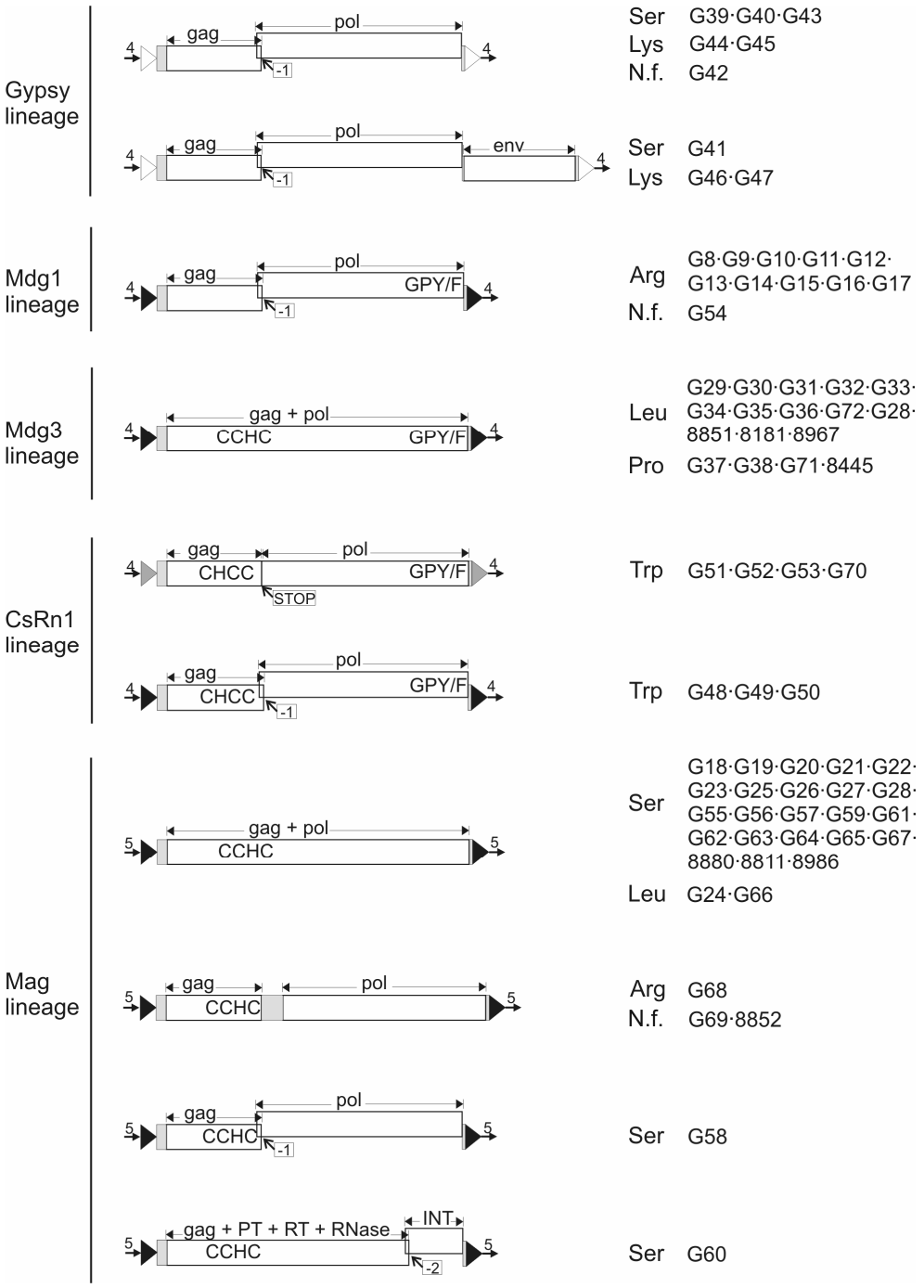



**Figura 15.** Principales características estructurales de los retrotransposones del grupo *Ty3/gypsy* de *A. gambiae*. Los rectángulos blancos representan ORFs, y los grises regiones no codificadoras. Los triángulos representan LTRs; los negros LTRs con terminaciones TG…CA; los grises LTRs con terminaciones TG…AA; y los blancos LTRs con terminaciones AG…YT. Se indican también los motivos CCHC y CHCC de *gag*, y el módulo GPY/F de la integrasa. En la parte derecha de la figura, se indican los tRNA complementarios al PBS, seguido de las familias que presentan dicha combinación de caracteres estructurales. "N.f." significa que el tRNA complementario al PBS no ha sido identificado. Los nombres de las familias han sido abreviados de la siguiente forma para ganar espacio: la palabra "GYPSY" ha sido abreviada a una "G", la terminación "_AG" en cada uno de los nombres ha sido suprimida. También se ha suprimido el prefijo "AAAB0100" en los nombres de las familias 8851, 8181, 8967, 8880, 8811, 8986 y 8852. Si dos ORFs están solapados, se indica en un cuadro el deslizamiento de la pauta en pares de bases. La palabra "stop" dentro de un cuadro indica que un codón de *stop* separa *gag* y *pol*. PT, proteasa; RT, reversotranscriptasa; RNase, ribonucleasa H; INT, integrasa.

De las relaciones filogenéticas reportadas en la figura 14, puede concluirse la siguiente representación para cada linaje del grupo *Ty3/gypsy* en *A. gambiae*: el linaje de *CsRn1* está conformado por ~10% (7/73) del total de familias, el linaje de *Gypsy* por ~26% (19/73), el linaje de *Mdg1* por ~15% (11/73), el linaje de *Mdg3* por ~22% (16/73) y el linaje de *Mag* por ~41% (30/73). Teniendo en cuenta la información reportada para *D. melanogaster* por Lerat, Rizzon & Biémont (2003) y por Kaminker et al. (2002), podemos concluir que tres de los seis linajes del grupo *Ty3/gypsy* en insectos (*Gypsy*, *Mdg1* y *Mdg3*) se encuentran bien representados en *A. gambiae* y *D. melanogaster* por varias familias; los linajes de *CsRn1* y *Gypsy*, bien representados en *A. gambiae*, están prácticamente ausentes en el genoma de *D. melanogaster*, resultando especialmente llamativo el caso del linaje *Mag*, para el que se han caracterizado en detalle hasta 30 familias en *A. gambiae* y, sin embargo, en *D. melanogaster* este linaje parece estar representado únicamente por algunas copias degeneradas; por último, no se ha encontrado huella alguna de elementos del linaje *Osvaldo* en el genoma de *A. gambiae*, aunque también en *D. melanogaster* el número de familias identificadas es muy bajo (Kapitonov & Jurka, 2003). En resumen, los datos obtenidos permiten, en general, concluir una mayor diversidad y abundancia de retrotransposones del grupo *Ty3/gypsy* en *A. gambiae* frente a *D. melanogaster*.

### IV.1.2.1. Diversidad estructural del grupo *Ty3/gypsy* en *A. gambiae*.

La caracterización molecular de los elementos representativos de cada una de las familias de retrotransposones del grupo *Ty3/gypsy* en *A. gambiae*, reveló una interesante diversidad estructural (figura 15):

Los elementos del linaje *Gypsy* presentan un deslizamiento de la pauta de lectura de -1pb entre *gag* y *pol*, y no presentan el dominio CCHC en el extremo C-terminal de *gag*. Además, las familias del linaje de *Gypsy* pueden presentar o no un tercer ORF *env*. Este ORF, que codifica presumiblemente para una envuelta, fue identificado en los elementos *GYPSY41_AG*, *GYPSY46_AG* y *GYPSY47_AG*. Teniendo en cuenta las relaciones filogenéticas mostradas en la figura 14, este hecho implica tres pérdidas independientes del ORF *env* durante la evolución de este linaje (ver ramas B, D y E en figura 14). Todos los elementos de las familias del linaje *Gypsy* en *A. gambiae* presentan LTRs que comienzan con los nucleótidos AG y terminan con YT, algo que diferencia a este linaje del resto de

linajes del grupo *Ty3/gypsy* y que también parece ocurrir en *D. melanogaster* (Kapitonov & Jurka, 2003). Las familias del linaje *Gypsy* pueden ser clasificados en dos grupos, en función de la presencia de un PBS complementario a $tRNA^{Ser}$ o a $tRNA^{Lys}$, tal y como ya fuera previamente reportado para *D. melanogaster* por (Terzian, Pelisson & Bucheton, 2001). La filogenia obtenida sugiere que el PBS complementario a $tRNA^{Lys}$ surgiría más tarde en la evolución del linaje *Gypsy* (cluster A en figura 14). Además, dentro del linaje *Gypsy* se han identificado dos casos claros de adquisición de una inserción preferencial. Así, los elementos de las familias *GYPSY41_AG*, *GYPSY42_AG* y *GYPSY43_AG* (cluster C en figura 14) presentan un TSD (*target site duplication*) ATAT, mientras que los elementos de las familias *GYPSY44_AG* y *GYPSY45_AG* (cluster B en figura 14) se insertan en lugares del genoma con el motivo C(G/T)CG. Se cree que la adquisición de mecanismos de inserción preferencial puede jugar un papel importante en la coevolución entre genoma hospedador y TE (SanMiguel et al., 1996; Voytas, 1996).

Los elementos del linaje *Mdg1*, al igual que el linaje *Gypsy*, presentan un deslizamiento de la pauta de lectura de -1pb entre *gag* y *pol*, y no presentan el dominio CCHC en el extremo C-terminal de *gag*. Sin embargo, los elementos de *Mdg1* presentan el módulo GPY/F de la integrasa, algo que los diferencia de *Gypsy*. Todas las familias del linaje de *Mdg1* presentan un PBS complementario a $tRNA^{Arg}$, excepto el elemento basal *GYPSY54_AG*, para el que no ha podido identificarse el PBS.

Los elementos del linaje de *Mdg3* presentan un único ORF que contiene *gag* y *pol*, el dominio CCHC de *gag* y el módulo GPY/F de la integrasa. A pesar de que las características estructurales de este linaje están claramente conservadas, el PBS ha evolucionado. Así, las familias del cluster F de la figura 14, presentan un PBS complementario a $tRNA^{Pro}$, en lugar del $tRNA^{Leu}$ que caracteriza a la mayoría de familias del linaje.

En el linaje de *CsRn1*, algunos de sus elementos pueden presentar un *stop codon* separando *gag* y *pol* (cluster G en figura 1), o bien ambos ORFs pueden presentar un deslizamiento de la pauta de lectura de -1pb. Una característica muy interesante, que diferencia a los elementos de este linaje del resto de linajes en *A. gambiae*, es un inusual dominio CHCC en el extremo C-terminal de *gag*, en lugar del dominio CCHC que presentan otros linajes. Además, estos elementos se diferencian por presentar un PBS complementario a $tRNA^{Trp}$. También presentan el módulo GPY/F de la integrasa.

Finalmente, resulta muy curioso que los elementos del cluster G (figura 14), que además se caracterizan por tener un deslizamiento de la pauta de lectura entre *gag* y *pol* de -1pb, presentan unos extremos terminales de sus LTRs diferentes al resto de los elementos del grupo *Ty3/gypsy* en *A. gambiae*: las LTRs comienzan con TG y terminan con AA, en lugar de las terminaciones TG…CA esperadas.

El linaje *Mag*, además de ser el más abundante, es el más diverso estructuralmente. Mayoritariamente, *gag* y *pol* suelen presentarse en un mismo ORF. Sin embargo, en algunas familias pueden (1) encontrarse separados por un área de repeticiones simples (cluster H en figura 14), (2) presentar un deslizamiento de la pauta de lectura de -1pb, o bien (3) la integrasa puede encontrarse en un ORF diferente que presenta un deslizamiento de la pauta de lectura de -2pb con respecto al otro gran ORF que contiene *gag* y el resto de dominios de *pol*. La mayoría de las familias presentan un PBS complementario a $tRNA^{Ser}$, sin embargo, también puede encontrarse un PBS complementario a $tRNA^{Leu}$ (cluster I en figura 14), o a $tRNA^{Arg}$ (elemento *GYPSY68_AG*). Otra característica que distingue a los elementos de este linaje sobre los demás es que tras su inserción en el genoma de *A. gambiae* generan un TSD de 5pb.

### IV.1.3. Evolución de los linajes de *Mdg3*, *Gypsy* y *Mdg1* en el genoma de *A. gambiae*.

Los tres linajes con mejor representación en los genomas de *A. gambiae* y *D. melanogaster* (*Mdg3*, *Gypsy* y *Mdg1*) muestran dos esquemas evolutivos diferentes (figura 14). Por un lado, para los linajes de *Mdg3* y *Gypsy* pueden observarse varias ramas que contienen familias pertenecientes a ambos genomas, indicativo de una clara diversificación de este linaje previa a la divergencia de los géneros *Anopheles* y *Drosophila*. Por otro lado, la topología de la filogenia muestra para el linaje de *Mdg1* la divergencia en dos grupos monofiléticos especie-específicos, con la excepción del elemento basal *GYPSY54_AG*. Este hecho, compatible con la transmisión vertical de las familias de dicho linaje, sugiere que la principal diversificación del linaje de *Mdg1* tuvo lugar paralelamente en ambos genomas tras la diversificación de los géneros *Drosophila* y *Anopheles*.

La evolución en mosaico con origen recombinacional toma una especial relevancia en la evolución del linaje de *Mdg1* en el genoma de *A. gambiae*. Costas, Valadé & Naveira

(2001) ya habían reportado anteriormente la importancia de estos procesos de recombinación en la evolución del linaje de *Mdg1* en *D. melanogaster*.

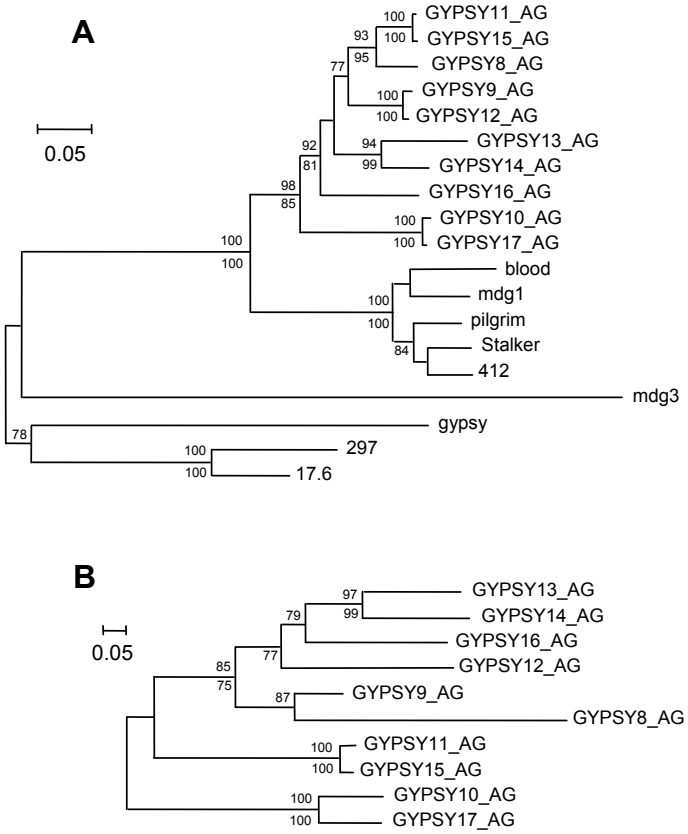



**Figura 16.** Relaciones filogenéticas entre las familias del linaje de *Mdg1* del genoma de *A. gambiae*. La topología de ambos árboles se obtuvo por el método NJ. Para las ramas interiores se muestran los valores *bootstrap* superiores al 70% (NJ encima de las ramas y MP debajo de las ramas). (A) Basado en el alineamiento de los dominios RT-RH-INT; (B) basado en el alineamiento de la región *gag*-like.

El análisis filogenético independiente de, por un lado, el alineamiento de los dominios RT-RH-INT y, por otro, los dominios conservados de la región *gag*, mostró para el linaje de *Mdg1* de *A. gambiae* una incongruencia significativa entre ambas filogenias que afecta a las familias *GYPSY8*, *GYPSY9* y *GYPSY12* (figura 16). En ambos árboles se repiten los tres *clusters GYPSY11*-*GYPSY15*, *GYPSY9*-*GYPSY12* y *GYPSY13*-*GYPSY14*. Sin embargo, atendiendo a la filogenia obtenida para la región *gag* (figura16-B), *GYPSY8* pasa a formar un cluster con *GYPSY9*, respaldado por un valor *bootstrap* del 87%, y *GYPSY12* se reagrupa con *GYPSY13*, *GYPSY14* y *GYPSY16*, mientras que en la filogenia

obtenida para la región *pol* (figura 16-A) *GYPSY9* y *GYPSY12* conforman un único cluster que presenta un valor *bootstrap* del 100%. La detección de estas discordancias llevó a la comparación dos a dos, mediante *dot plots*, de los tres elementos implicados (Tubío, Costas & Naveira, 2004), permitiendo confirmar que el origen de estas diferencias se debe al efecto de la recombinación entre los elementos *GYPSY8* y *GYPSY9* y entre los elementos *GYPSY12* y *GYPSY16*, posiblemente tras el empaquetamiento de dos moléculas de RNA en una misma VLP (McDonald, 1993; Mikkelsen & Pedersen, 2000).

### IV.1.4. Evolución del linaje de *Osvaldo*.

Resulta especialmente llamativa la ausencia de elementos *Osvaldo*-like en el genoma de *A. gambiae*, dado que se ha podido constatar una buena representación de este linaje en los genomas de los mosquitos *Aedes aegypti* (Nene et al., 2007) y *Culex pipiens* (manuscrito en preparación). A pesar de la existencia de huellas de actividad de este linaje en el genoma de *D. melanogaster* (Kapitonov & Jurka, 2003), no se han podido encontrar ni siquiera rastros de una actividad ancestral en el genoma de *A. gambiae*.

Una posible explicación para la virtual ausencia de elementos *Osvaldo*-like podría tener que ver con la presencia de LTRs muy largas, siempre mayores a mil pares de bases, que caracteriza a los elementos de este linaje. Esta característica ha sido observada en los genomas de *Aedes aegypti* (Nene et al., 2007), *Culex pipiens* (manuscrito en preparación) y de *Drosophila melanogaster*. De hecho, el elemento *Gypsy12* de *D. melanogaster*, perteneciente al linaje de *Osvaldo*, es el que presenta las LTRs de mayor tamaño dentro de los retrotransposones de esta especie, alcanzando una longitud aproximada de 2300pb (Kapitonov & Jurka, 2003). Una característica común a casi todos los linajes del grupo *Ty3/gypsy* en *A. gambiae* es que las LTRs que presentan sus elementos no son demasiado largas, estando la mayoría de ellas por debajo de las 400pb. La única excepción a este patrón es el elemento *GYPSY69_AG* del linaje *Mag* (con 441pb) y 10 elementos del linaje de *Mdg1*, con tamaños comprendidos entre 404-982pb. Esta característica general de los retrotransposones de *A. gambiae* (con la excepción del linaje de *Mdg1*) podría ser el resultado de una mayor eficacia de la selección natural actuando en contra de aquellas familias de elementos con LTRs largas, dado que serían mucho más propensas a estar implicadas en fenómenos de recombinación ectópica con potenciales efectos deletéreos. Si

se considera que la recombinación ocurre con igual frecuencia a lo largo de toda la longitud de las LTRs, entonces los retrotransposones con mayores LTRs tenderían a recombinar con mayor frecuencia que aquellos con LTRs menores. Una demostración a pequeña escala de este efecto podría obtenerse estudiando la frecuencia con la que se producen LTRs solitarias, por recombinación entre ambas LTRs de una misma inserción, en función del tamaño de las LTRs de una familia de retrotransposones. En este sentido, Vitte & Panaud (2003) compararon en el arroz tres familias de elementos con diferentes tamaños de LTR, encontrando que la proporción de LTR-solitarias parece incrementarse con el mayor tamaño de las LTRs, aunque los autores admiten que el tamaño de muestra es bajo. En el caso de *A. gambiae*, los elementos *Ty3/gypsy* con mayores LTRs corresponden al linaje de *Mdg1*, precisamente el linaje con mayor proporción de LTRs solitarias, lo que soportaría esta hipótesis (tabla 1).

**Tabla 1**. Proporción de Solo-LTRs *vs* provirales en los linajes del grupo *Ty3/gypsy* de *A. gambiae*.

| Lineage | LTR:proviral[a] | LTR size[b] | Range[c] | $N_{fam}$[d] |
|---|---|---|---|---|
| Gypsy | 0.8:1.0 (35:44) | 258.11 | 167-353 | 9 |
| Mdg1 | 2.0:1.0 (158:79) | 740.20 | 404-982 | 10 |
| Mdg3 | 0.6:1.0 (48:80) | 223.25 | 161-294 | 12 |
| CsRn1 | 0.38:1.0 (24:64) | 321.71 | 146-393 | 7 |
| Mag | 0.4:1.0 (64:158) | 229.93 | 108-441 | 15 |

[a] Proporción de Solo-LTRs frente a copias provirales. La categoría "provirales" incluye las categorías "activas" e "inactivas". Entre paréntesis se muestra el número total de copias contabilizadas de Solo-LTRs frente a copias provirales.
[b] Tamaño medio de las LTRs del linaje.
[c] Rango del total de tamaños de las LTRs del linaje.
[d] Número total de familias analizadas del linaje.

### IV.1.5. Evolución de los linajes de *CsRn1* y *Mag*.

Los otros dos principales linajes de insectos (*CsRn1* y *Mag*) están prácticamente ausentes del genoma de *D. melanogaster*, aunque bien representados en el genoma de *A. gambiae*, representando un caso extremo el correspondiente al linaje de *Mag*, ya que resulta curioso que el linaje de *Mag* represente aproximadamente el 40% del componente

*Ty3/gypsy* del genoma de este mosquito, mientras que en *D. melanogaster* se encuentre prácticamente ausente. Este hecho puede resultar mucho más curioso todavía si se tiene en cuenta que el linaje de *Mag* parece estar representado en el Reino Animal no solo en insectos, sino también en vertebrados (Volff et al., 2001). Es más, en los genomas de otros mosquitos, como *Aedes aegypti* (Nene et al., 2007) o *Culex pipiens* (manuscrito en preparación), llega a conformar, nuevamente, el componente mayoritario entre los retrotransposones del grupo *Ty3/gypsy*. El fracaso o éxito evolutivo de una familia de TEs en los genomas es bastante complejo, atendiendo tanto a características propias de la biología de las mismas, como a características propias del genoma hospedador y a la historia evolutiva propia de las especies (Charlesworth, Sniegowski & Stephan, 1994).

### IV.1.6. Tasa de renovación de los retrotransposones del genoma de *A. gambiae.*

La tabla 2 muestra la clasificación del total de copias de 63 familias del grupo *Ty3/gypsy* en *A. gambiae*. Estas familias presentan una importante proporción de copias presumiblemente inactivas, la mayoría de ellas degeneradas. Al menos un 41% del total de copias analizadas (264/642) corresponden a este grupo, mientras que solamente un 13% de las copias analizadas (85/642) corresponden a la categoría de presumiblemente activas. Es más, la proporción de copias inactivas aumenta a un 69% (440/642), si se considera el número de LTRs solitarias en la categoría de copias inactivas. Además, la divergencia entre copias inactivas comparadas dos a dos (tabla 3) es significativamente mayor que la divergencia entre copias presumiblemente activas ($t=-6.319$; $P<0.001$). Estos datos revelan, para *A. gambiae*, un patrón muy diferente al conocido para *D. melanogaster*.

Lerat, Rizzon & Biémont (2003) revelaron que las familias de TEs en el genoma de *D. melanogaster* se caracterizan por una elevada homología de secuencia y, por tanto, una carencia de elementos divergentes. En este sentido, Kaminker et al. (2002) ya reportaran resultados equivalentes, tras observar que todas las familias de retrotransposones con LTRs de *D. melanogaster* presentan una divergencia media entre sus copias inferior al 1%. Es más, ~45% de las copias reportadas por Kaminker (304/682) tenían estructura completa. Lerat, Rizzon & Biémont (2003) sugerían que la ausencia de copias divergentes del grupo de los retrotransposones en *D. melanogaster* podría ser debida a una elevada tasa de renovación, que consistiría en la rápida eliminación de copias inactivas. Un reciente

estudio llevado a cabo por Bergman & Bensasson (2007) sugiere que las estimas de las edades de los retrotransposones realizadas por Kaminker et al. (2002) y por Lerat, Rizzon & Biémont (2003), en base a los datos de divergencia (*average pairwise distance*), estarían sesgados por influencia de la constricción selectiva a la que estarían sometidos los elementos. Sin embargo, los autores confirman que los retrotransposones con LTRs están representados por inserciones muy recientes. De hecho, según Bergman & Bensasson (2007), ~19% de todas las inserciones de los retrotransposones con LTRs se habrían insertado hace no más de 92600 años. Estos resultados confirman los previamente reportados por Bowen & McDonald (2001) para *D. melanogaster*, que concluían edades de inserción recientes basándose en la divergencia entre ambas LTRs de la misma copia. Además, los resultados de Bergman & Bensasson (2007) indican que la mayoría de las inserciones de los retrotransposones con LTRs de *D. melanogaster* se habrían insertado hace aproximadamente 16 mil años. Curiosamente, las estimas obtenidas para la inserción de retrotransposones sin LTRs en *D. melanogaster* son mucho mayores (en torno a 50 veces más que para los retrotransposones con LTRs), por lo que muchas de éstas se habrían insertado en poblaciones ancestrales africanas de *D. melanogaster*.

Sobre los mecanismos de eliminación de retrotransposones el más conocido es la formación de LTRs solitarias, a través de la recombinación entre ambas LTRs de una misma copia. En este sentido, uno de los casos mejor estudiados es el que corresponde al genoma de *Saccharomyces cerevisiae* (Jordan & McDonald, 1998; Kim et al., 1998; Jordan & McDonald, 1999), donde las familias de retrotransposones con LTRs están compuestas por copias completas altamente homogéneas y por una elevada cantidad de LTRs solitarias. La carencia de copias degeneradas inactivas se debe, en el caso de *Saccharomyces*, a una elevada tasa de formación de LTRs, persistiendo en el genoma únicamente aquellas copias completas con transposición más reciente. Sin embargo, este mecanismo *per se* no es suficiente para explicar la carencia de copias degeneradas inactivas de retrotransposones con LTRs en el genoma de *D. melanogaster* porque, mientras que en *Saccharomyces* la proporción de LTRs solitarias con respecto al número total de copias es del 85% (Kim et al., 1998), el número de LTRs solitarias encontradas en *D. melanogaster* es mucho menor. Tal y como indican Lerat, Rizzon y Biémont (2003), el mayor número de dichas LTRs solitarias corresponde a la familia *roo/B104* con un 18% (22 LTRs solitarias de un total de 125 copias) y el resto de elementos presentan números

mucho menores: la familia *tinker* presenta un 11% (1 LTR solitaria de un total de 9 copias), la familia *stalker* un 9% (1/11), la familia *412* un 3.3% (1/30), la familia *mdg1* un 4.7% (1/21) y la familia *297* un 7.41% (4/54). Los datos previamente reportados por Kaminker et al. (2002) están en la misma línea, ya que solamente identificaron 58 LTRs solitarias entre todas las familias de retrotransposones de *D. melanogaster* (14 de éstas pertenecían a la familia *roo*).

Probablemente, la menor tasa de renovación de retrotransposones observada para *A. gambiae* requiera una explicación compleja, que comprenda diversos aspectos biológicos propios del genoma y de los TEs pero, por supuesto, donde también debe ser tenida en cuenta la historia evolutiva de cada una de las líneas que siguió a la divergencia de ambos géneros y que dio lugar a ambas especies. La persistencia de una proporción tan elevada de copias fragmentadas y divergentes en la mayoría de las familias de retrotransposones del grupo *Ty3/gypsy* en el genoma de *A. gambiae*, indica que muchas de las inserciones que residen en este genoma son antiguas. Dado el carácter deletéreo derivado de la actividad de los TEs en los genomas, es probable que muchas de esas inserciones se encuentren fijadas, lo que explicaría que no hayan sido eliminadas por acción de la selección. Con la intención de obtener unos primeros datos acerca de la posibilidad de esta hipótesis, se ha procedido a la búsqueda de inserciones fijadas, estimando las tasas de ocupación de algunas inserciones antiguas en el genoma de este mosquito.

**Tabla 2.** Número total de copias de las familias del grupo *Ty3/gypsy* en el genoma *Anopheles gambiae* clasificadas en función de su condición de actividad.

| Lineage | Family | Total | Active[a] | Inactive | Sol-LTR |
|---|---|---|---|---|---|
| *Mag* | G18 | 22 | 6(6) | 10 | 3 |
| *Mag* | G19 | 7 | 2(2) | 5 | 0 |
| *Mag* | G20 | 10 | 1(1) | 5 | 3 |
| *Mag* | G21 | 23 | 2(2) | 12 | 7 |
| *Mag* | G22 | 4 | 2(1) | 2 | 0 |
| *Mag* | G23 | 4 | 1(0) | 1 | 1 |
| *Mag* | G24 | 6 | 0(0) | 4 | 1 |
| *Mag* | G25 | 9 | 1(1) | 5 | 2 |
| *Mag* | G26 | 6 | 0(0) | 4 | 0 |
| *Mag* | G27 | 5 | 1(0) | 2 | 1 |
| *Mag* | G28 | 4 | 0(0) | 1 | 2 |
| *Mag* | G55 | 12 | 0(0) | 3 | 6 |
| *Mag* | G56 | 7 | 1(1) | 3 | 1 |
| *Mag* | G57 | 15 | 2(2) | 5 | 6 |
| *Mag* | G58 | 14 | 2(1) | 2 | 7 |
| *Mag* | G59 | 8 | 0(0) | 4 | 2 |
| *Mag* | G60 | 8 | 2(2) | 3 | 0 |
| *Mag* | G61 | 7 | 1(1) | 2 | 2 |
| *Mag* | G62 | 10 | 2(2) | 4 | 3 |
| *Mag* | G63 | 14 | 1(1) | 8 | 4 |
| *Mag* | G64 | 7 | 1(1) | 3 | 1 |
| *Mag* | G65 | 5 | 2(2) | 3 | 0 |
| *Mag* | G66 | 6 | 1(0) | 4 | 0 |
| *Mag* | G67 | 11 | 0(0) | 6 | 3 |
| *Mag* | G68 | 6 | 2(1) | 3 | 0 |
| *Mag* | G69 | 10 | 1(1) | 5 | 1 |
| *Mag* | 8880 | 11 | 1(0) | 5 | 3 |
| *Mag* | 8811 | 7 | 1(1) | 3 | 2 |
| *Mag* | 8986 | 4 | 1(1) | 2 | 1 |
| *Mag* | 8852 | 4 | 1(1) | 1 | 2 |
| *Mdg3* | G29 | 7 | 4(4) | 2 | 1 |
| *Mdg3* | G30 | 18 | 3(2) | 3 | 6 |
| *Mdg3* | G31 | 11 | 2(1) | 2 | 6 |
| *Mdg3* | G32 | 19 | 7(7) | 4 | 2 |
| *Mdg3* | G33 | 6 | 0(0) | 3 | 3 |
| *Mdg3* | G34 | 15 | 1(1) | 7 | 3 |
| *Mdg3* | G35 | 18 | 4(3) | 4 | 3 |
| *Mdg3* | G36 | 7 | 1(1) | 2 | 2 |
| *Mdg3* | G37 | 8 | 1(1) | 3 | 3 |
| *Mdg3* | G38 | 11 | 0(0) | 4 | 2 |
| *Mdg3* | G71 | 15 | 0(0) | 5 | 8 |
| *Mdg3* | G72 | 7 | 0(0) | 4 | 1 |
| *Mdg3* | 8851 | 3 | 1(1) | 1 | 1 |
| *Mdg3* | 8181 | 5 | 1(0) | 2 | 2 |
| *Mdg3* | 8967 | 10 | 1(0) | 4 | 5 |
| *Mdg3* | 8445 | 5 | 1(1) | 3 | 0 |
| *Gypsy* | G39 | 23 | 1(1) | 6 | 11 |
| *Gypsy* | G40 | 6 | 0(0) | 3 | 2 |
| *Gypsy* | G41 | 12 | 0(0) | 6 | 6 |
| *Gypsy* | G42 | 7 | 1(0) | 2 | 1 |
| *Gypsy* | G43 | 4 | 0(0) | 4 | 0 |
| *Gypsy* | G44 | 6 | 1(1) | 3 | 1 |
| *Gypsy* | G45 | 9 | 1(1) | 2 | 4 |
| *Gypsy* | G46 | 14 | 2(1) | 5 | 6 |
| *Gypsy* | G47 | 11 | 0(0) | 7 | 4 |

| | | | | | |
|---|---|---|---|---|---|
| *CsRn1* | G48 | 7 | 1(1) | 2 | 2 |
| *Gypsy* | G49 | 4 | 1(1) | 2 | 0 |
| *Gypsy* | G50 | 28 | 5(5) | 14 | 7 |
| *Gypsy* | G51 | 25 | 5(5) | 6 | 3 |
| *Gypsy* | G52 | 23 | 1(1) | 14 | 6 |
| *Gypsy* | G53 | 8 | 0(0) | 5 | 2 |
| *Gypsy* | G70 | 16 | 0(0) | 8 | 4 |
| *Mdg1* | G54 | 8 | 1(1) | 2 | 5 |
| TOTAL | 631 | 642 | 85(71) | 264 | 176 |

[a] Número de copias activas. Entre paréntesis se indica el número de copias activas que presentan LTRs idénticas.

**Tabla 3**. Media de la identidad entre los elementos activos y entre los inactivos comparados dos a dos, para las familias del grupo *Ty3/gypsy* en *Anopheles gambiae*.

| **Family**[a] | **Id$_{act}$**[b] | **Range**[c] | **N**[d] | **Id$_{in}$**[e] | **Range**[c] | **N**[d] |
|---|---|---|---|---|---|---|
| G18 | 99.93 | 99.87-99.98 | 15 | 94.94 | 91.16-99.58 | 35 |
| G19 | 99.91 | – | 1 | 96.93 | 94.69-99.91 | 9 |
| G21 | 99.49 | – | 1 | 94.20 | 90.56-98.26 | 32 |
| G22 | 99.89 | – | 1 | 99.94 | – | 1 |
| G57 | 99.29 | – | 1 | 95.40 | 91.75-98.69 | 9 |
| G60 | 99.98 | – | 1 | 94.79 | 93.39-96.23 | 3 |
| G62 | 99.89 | – | 1 | 91.21 | 88.14-95.78 | 5 |
| G65 | 99.92 | – | 1 | 96.80 | 95.79-98.33 | 5 |
| G68 | 99.85 | – | 1 | 96.65 | 95.10-99.63 | 3 |
| G29 | 99.90 | 99.86-99.94 | 6 | 99.78 | – | 1 |
| G30 | 99.93 | 99.92-99.94 | 3 | 96.16 | – | 1 |
| G31 | 99.92 | – | 1 | 93.92 | – | 1 |
| G32 | 99.76 | 99.27-100 | 21 | 97.55 | 96.81-98.75 | 6 |
| G35 | 99.17 | 99.93-99.32 | 6 | 98.70 | 98.04-99.72 | 6 |
| G46 | 99.41 | – | 1 | 96.30 | 92.14-99.20 | 5 |
| G50 | 99.98 | 99.95-100 | 10 | 95.70 | 91.67-99.99 | 40 |
| G51 | 99.12 | 99.81-99.24 | 10 | 97.57 | 97.01-97.93 | 10 |

[a] Nombres abreviados de las familias.

[b] Media de la identidad de las comparaciones dos a dos de los elementos activos de la familia.

[c] Rango del total de valores de identidad observados.

[d] Número total de comparaciones.

[e] Media de la identidad de las comparaciones dos a dos de los elementos inactivos de la familia (se excluyen Solo-LTRs).

### IV.1.7. Tasas de ocupación de elementos transponibles en *Anopheles gambiae*.

El estudio llevado a cabo comprendía el análisis de 163 muestras de ADN de mosquitos de la especie *A. gambiae*. De éstas, 98 muestras correspondían a mosquitos recolectados en la localidad de Goundry (Burkina Faso) al Oeste de África, mientras que las restantes 65 muestras correspondían a mosquitos recolectados en dos localidades de Kenia, al Este de África. Estos dos lugares de Kenia, conocidos como Asembo Bay (a partir de ahora en el texto, Asembo) y Jego, se encuentran separadas por el Valle del Rift, que supone una barrera al flujo génico en la especie. Asembo se encuentra localizada en el Oeste de Kenya (y, por tanto, al Oeste del Valle del Rift), mientras que Jego se encuentra ubicada en el Este de Kenya, cerca de la costa. Todos los individuos analizados en Kenia corresponden a la forma molecular *S* de *A. gambiae*, mientras que en la localidad de Burkina Faso, donde ambas formas moleculares se dan en simpatría, se han analizado 50 muestras de individuos pertenecientes a la forma molecular *M* y 48 pertenecientes a la forma molecular *S*. El diseño de esta parte del experimento pretendía tener en cuenta toda la variabilidad conocida que afecta a la estructuración de la especie *A. gambiae* a lo largo del continente africano. Además, teniendo en cuenta la probable situación filogenética basal de la especie *A. arabiensis* en el Complejo de *A. gambiae*, también se han analizado 20 muestras de ADN de mosquitos de esta especie. Estas muestras correspondían a 10 mosquitos recolectados en Goundry (Burkina Faso) y a 10 mosquitos recolectados en Asembo (Kenia).

Dado que el objetivo prioritario del estudio era encontrar inserciones fijadas en *A. gambiae*, llevamos a cabo una selección de *loci* candidatos, tomando como principales criterios dos: (1) deberían ser loci antiguos, y (2) deberían ser de tamaño pequeño. El motivo del primer criterio es lógico, si tenemos en cuenta que cuanto más antiguo sea un locus mayor será la probabilidad de que éste se encuentre fijado. El segundo criterio se tomó, sencillamente, por razones metodológicas, dado que se buscaban *loci* de entre 150-700pb para poder ser amplificados fácilmente por PCR estándar. Atendiendo a estas dos premisas, se filtraron 642 TEs de una base de datos previamente elaborada por nuestro grupo de investigación (Tubío, Naveira & Costas, 2005), de la que se seleccionaron los 10 *loci* que mejor se ajustaban a los criterios de búsqueda. Finalmente, tras varias pruebas de optimización de las PCRs sobre estos 10 *loci* candidatos, se escogieron aquellos 3 *loci* con mejor resultados en la amplificación (tabla 4).

**Tabla 4.** Características de los *loci* seleccionados en *A. gambiae* para el estudio de las tasas de ocupación.

| Loci | Family[a] | Scaffold[b] | Chr[c] | Position[d] | Size[e] | Div[f] |
|---|---|---|---|---|---|---|
| *Locus1* | GYPSY31_AG | AAAB01008980 | 3R | 7409466-7409287 | 179 | 0.6 |
| *Locus2* | 8967 | AAAB01008881 | NP | 208366-208554 | 188 | 0.6 |
| *Locus3* | GYPSY57_AG | AAAB01008960 | 2L | 5250968-5251267 | 299 | 0.6 |

[a] Nombre de la familia a la que pertenece el locus estudiado (Tubío, Naveira & Costas, 2005).
[b] *Scaffold* dentro del genoma secuenciado de la cepa PEST de *A. gambiae* donde se localiza el *locus*.
[c] Cromosoma donde se localiza el *locus*.
[d] Posición del locus dentro del *scaffold*.
[e] Tamaño del locus.
[d] Divergencia del locus estudiado con respecto a la secuencia consenso para la familia.

### IV.1.7.1. Elementos transponibles fijados en el genoma de *A. gambiae*.

El genotipado de los tres *loci* estudiados permitió estimar una tasa de ocupación del 100%, habiendo sido detectados en todos los individuos de las poblaciones y subpoblaciones analizadas. Es más, esta observación fue verificada mediante la secuenciación de al menos un amplicón en cada una de las poblaciones, que confirmó la presencia de los tres *loci* (ver alineamientos en páginas 143-145). La secuencia de los límites (5´ y 3´) de cada uno de los *loci* resultó ser más de un 99% idéntico a la secuencia publicada del genoma de la cepa PEST de *A. gambiae* (Holt et al., 2002). Además, la comparación de la secuencia total obtenida para los tres loci en las poblaciones y subpoblaciones estudiadas es consistente con la idea de que éstos se encuentran fijados en el genoma del mosquito *A. gambiae*. Es más, la secuenciación del *locus1* y del *locus3* en la especie *A. arabiensis* (que representa la especie basal dentro del Complejo de *Anopheles gambiae*) reveló que ambos *loci* también se encuentran presentes en el genoma de esta especie.

La fijación de estas tres inserciones antiguas en el genoma de *A. gambiae* es consistente con nuestra hipótesis de partida (Tubío, Costas & Naveira, 2004) que suponía altas tasas de ocupación en aquellas inserciones más antiguas que no han sido eliminadas por la acción de la selección. Estos resultados, compatibles con las suposiciones que establecen los modelos derivados de la teoría sobre la evolución del ADN parasítico, pueden ser interpretados principalmente como (1) una relajación de la acción purificadora

de la selección natural, o (2) la acción de la deriva genética como consecuencia de la reducción en el tamaño efectivo de población durante la historia evolutiva de la especie.

El genoma de *A. gambiae* presenta un tamaño total de 278Mb, mientras que el de *D. melanogaster* presenta 168Mb, y los exones suponen ~20Mb y ~31Mb (Holt et al., 2002), respectivamente en ambos genomas. Así, la densidad génica (entendida como número de bases ocupadas por secuencias codificadoras partido por el tamaño total del genoma) es significativamente menor en el genoma de *A. gambiae* frente al de *D. melanogaster* (chi-cuadrado=13.106; P<<0.005). Compatible con la primera de ambas interpretaciones propuestas, la menor densidad génica del genoma de *A. gambiae* con respecto al de *D. melanogaster*, más compacto, podría suponer una actuación menos estricta de la selección natural sobre el genoma de *Anopheles* en comparación con el de *Drosophila* y, como consecuencia de esta relajación de la acción purificadora de la selección natural, dichas copias podrían llegar a fijarse. La reciente secuenciación del genoma del mosquito *Ae. aegypti* supone una extraordinaria oportunidad para testar la hipótesis de la densidad génica. El genoma de este mosquito (1.3Gb) es más de 4.5 veces mayor que el de *A. gambiae*. Se ha estimado que los exones suponen ~26Mb del genoma de Aedes (Nene et al., 2007), de manera que la densidad génica en este mosquito es significativamente menor con respecto a la de *A. gambiae* (chi-cuadrado=24.188; P<<0.005). Con la idea de comparar las tasas de renovación de retrotransposones en ambos genomas, se ha llevado a cabo un estudio sobre la representación de estos TEs en el genoma de *Ae. aegypti*.

## IV.2. Retrotransposones con LTRs del grupo *Ty3/gypsy* en *Ae. aegypti*.

La búsqueda de TEs del grupo *Ty3/gypsy* en *Ae. aegypti*, empleando los mismos criterios definidos para la búsqueda en el genoma de *A. gambiae*, nos ha llevado a identificar hasta 184 familias bien representadas en el genoma de este mosquito (Nene et al., 2007). La filogenia resultante del alineamiento de los dominios RT-RH-INT permitió revelar que los seis principales linajes del grupo *Ty3/gypsy* de insectos se encuentran ampliamente representados en el genoma de este mosquito (figura 17), a diferencia del patrón anteriormente indicado para *A. gambiae* y *D. melanogaster*, distribuyéndose de la siguiente manera: ~16% (29/184) de las familias pertenecen al linaje *Osvaldo*; ~9%

(16/184) de familias al linaje *CsRn1*; ~15% (28/184) de familias al linaje *Mdg1*; ~13% (24/184) de familias al linaje *Mdg3*; ~35% (64/184) de familias al linaje *Mag* y ~13% (23/184) de familias al linaje *Gypsy*.

Sobre estos datos preliminares, destaca la extraordinaria abundancia del linaje *Osvaldo*, en detrimento de la abundancia de los linajes de *Mdg3*, de *Mag* y, sobre todo, de *Gypsy*. Este patrón contrasta con el reportado para el genoma de *A. gambiae*, donde no se han identificado elementos que pudieran ser asignados al linaje *Osvaldo*. En cuanto a los otros dos grandes linajes (*CsRn1* y *Mdg1*), no existen diferencias para la abundancia relativa de éstos en ambos genomas.

Atendiendo a la representación a nivel de Familia (tabla 5), la tasa de renovación de los retroelementos muestra un patrón general que se aproxima más al exhibido por el genoma de *A. gambiae* que al del genoma de *D. melanogaster*: solamente un 17% (127/760) del total de copias analizadas corresponde a inserciones activas, mientras que los elementos inactivos y las LTRs solitarias conforman, respectivamente, un 53% (404/760) y un 19% (144/760) del total de inserciones analizadas. La comparación de los patrones exhibidos por ambos genomas no muestra diferencias significativas para copias activas e inactivas (chi-cuadrado=0.022, P=0.882), incluso si se consideran en la categoría de inactivas las LTRs solitarias (chi-cuadrado=1.398, P=0.237). El componente de copias que no han podido ser clasificadas (85/760 para *Ae. Aegypti* y 117/ 642 para *A. gambiae*) tampoco varía este resultado cuando el número de copias no clasificadas es integrado en la categoría de inactivas provirales (chi-cuadrado=0.951, P=0.329) o en la categoría de inactivas provirales+LTRs solitarias (chi-cuadrado=3.266, P=0.071). Es más, la comparación entre los valores de *pairwise identity* obtenidos entre inserciones activas y entre inserciones inactivas en cada familia del linaje de *CsRn1* (tabla 6) muestra que son significativamente distintos (test-t student=-5.803, P<0.001). Al igual que para *A. gambiae*, la media de los valores *pairwise identity* entre las copias activas siempre supera el 99.0%. Con respecto a las copias inactivas, la media de los valores *pairwise identity* para cada una de las familias estudiadas (90.2%-98.0%) están dentro de los rangos obtenidos para genoma de *A. gambiae*.

Así, parece que el genoma de *Ae. aegypti* no alberga una cantidad significativamente mayor de copias activas o inactivas con respecto al genoma de *A. gambiae*, a pesar de que el genoma de *Ae. aegypti* presenta una densidad génica

significativamente menor que el de *A. gambiae*. Por tanto, la propuesta que sugiere que las diferencias observadas con respecto a la tasa de renovación de retrotransposones entre los genomas de *D. melanogaster* y de *A. gambiae* pudiera deberse a la diferencia en la densidad génica de ambos genomas, no se ve respaldada por extensión de los resultados observados para *A. gambiae* y *Ae. aegypti*. Es más, los valores *pairwise identity* obtenidos para las copias activas e inactivas de cada familia en los genomas de *A. gambiae* y *Ae. aegypti* (tabla3 y tabla 6) no difieren significativamente al comparar ambos genomas entre si (activas: t=-1.292, P=0.21; inactivas: t=-2.356, P=0.028). Estos datos indican, nuevamente, una tasa de renovación para los retrotransposones de *Ae. aegypti* que no difiere a la observada para *A. gambiae* y que, sin embargo, contrasta con la conocida para *D. melanogaster*.

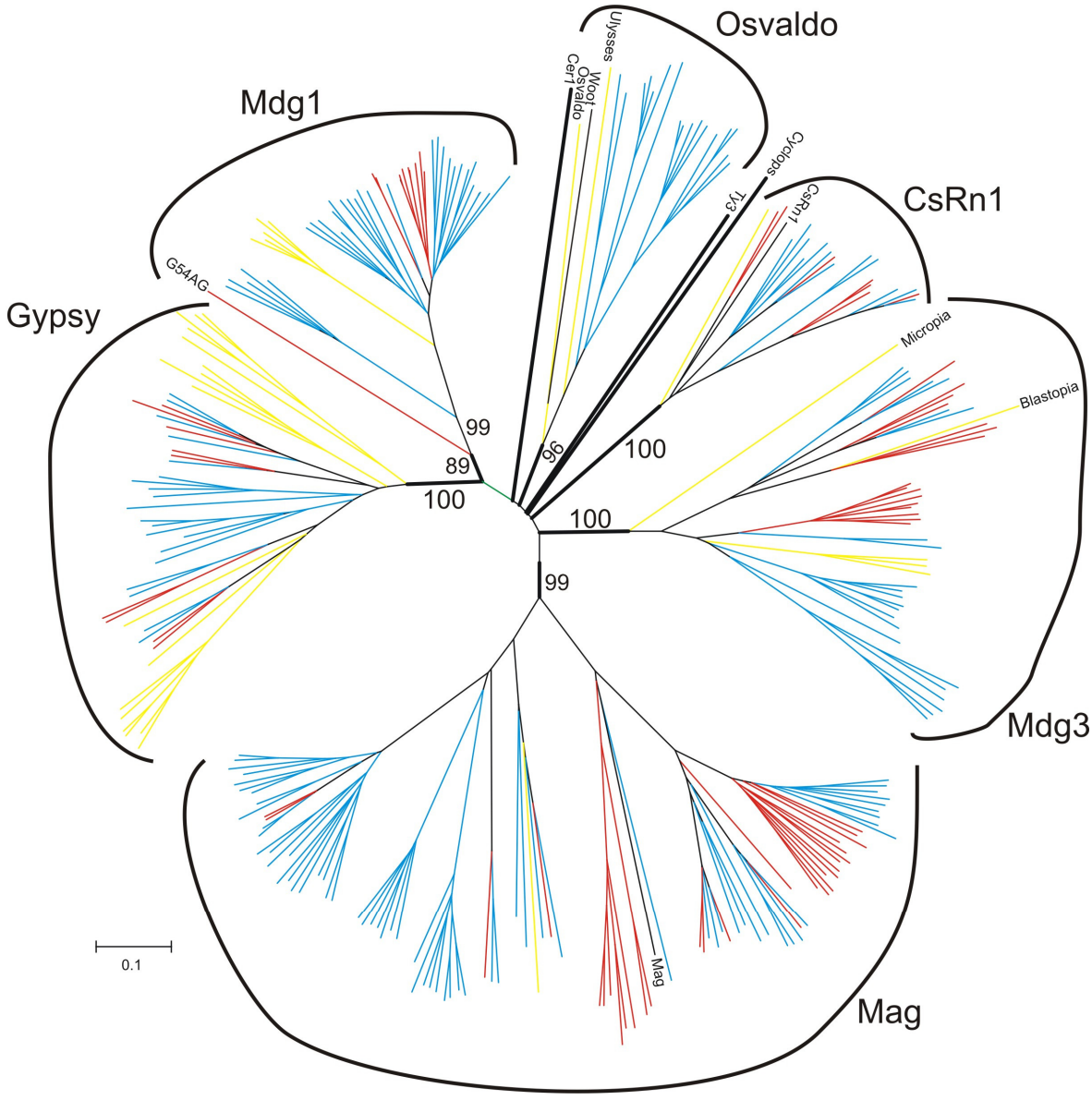



**Figura 17**. Topología general del árbol filogenético, generado por el método N-J, tras el alineamiento de los dominios RT-RH-INT de las 173 familias del grupo *Ty3/gypsy* de *Ae. aegypti* (Nene et al., 2007) (ramas azules), las 63 familias bien caracterizadas del grupo *Ty3/gypsy* de *A. gambiae* (Tubío, Naveira & Costas, 2005) (ramas rojas) y todas las familias conocidas del grupo *Ty3/gypsy* de *D. melanogaster* (ramas amarillas). También se han añadido a este alineamiento los elementos representativos de otras especies de cada unos de los linajes del grupo *Ty3/gypsy* que no tienen representación en *D. melanogaster* y *A. gambiae* (ramas terminales negras). Los valores *bootstrap* solo se indican en las ramas basales y fueron obtenidos por NJ (1000 réplicas).

**Tabla 5.** Número total de copias de las familias del grupo *Ty3/gypsy* en el genoma *Aedes aegypti* clasificadas en función de su condición de actividad.

| Lineage | Family | Total | Active[a] | Inactive | Solo-LTRs |
|---|---|---|---|---|---|
| *CsRn1* | *Ele1* | 9 | 4(4) | 2 | 0 |
| *CsRn1* | *Ele2* | 21 | 0 | 15 | 5 |
| *CsRn1* | *Ele3* | 18 | 5(3) | 10 | 3 |
| *CsRn1* | *Ele4* | 9 | 2(1) | 4 | 2 |
| *CsRn1* | *Ele5* | 5 | 2(1) | 3 | 0 |
| *CsRn1* | *Ele6* | 9 | 1(1) | 5 | 3 |
| *CsRn1* | *Ele7* | 14 | 1(0) | 6 | 7 |
| *CsRn1* | *Ele8* | 15 | 2(1) | 13 | 0 |
| *CsRn1* | *Ele9* | 5 | 1(1) | 3 | 1 |
| *CsRn1* | *Ele10* | 12 | 2(0) | 5 | 4 |
| *CsRn1* | *Ele11* | 7 | 2(0) | 4 | 1 |
| *CsRn1* | *Ele12* | 8 | 0 | 5 | 1 |
| *CsRn1* | *Ele109* | 7 | 1(0) | 5 | 1 |
| *CsRn1* | *Ele110* | 14 | 0 | 6 | 8 |
| *CsRn1* | *Ele111* | 7 | 0 | 4 | 1 |
| *CsRn1* | *Ele112* | 7 | 1(0) | 5 | 1 |
| *Gypsy* | *Ele51* | 5 | 1(1) | 4 | 0 |
| *Gypsy* | *Ele52* | 12 | 0 | 9 | 2 |
| *Gypsy* | *Ele53* | 18 | 5(1) | 11 | 0 |
| *Gypsy* | *Ele54* | 11 | 5(1) | 6 | 0 |
| *Gypsy* | *Ele55* | 37 | 1(0) | 25 | 9 |
| *Gypsy* | *Ele56* | 8 | 2(2) | 5 | 1 |
| *Gypsy* | *Ele122* | 71 | 7(3) | 42 | 14 |
| *Gypsy* | *Ele123* | 11 | 2(0) | 5 | 4 |
| *Gypsy* | *Ele124* | 9 | 0 | 6 | 1 |
| *Gypsy* | *Ele130* | 5 | 0 | 4 | 1 |
| *Mdg1* | *Ele57* | 18 | 7(3) | 5 | 4 |
| *Mdg1* | *Ele58* | 57 | 11(4) | 23 | 6 |
| *Mdg1* | *Ele59* | 44 | 4(0) | 24 | 8 |
| *Mdg1* | *Ele60* | 24 | 3(0) | 10 | 6 |
| *Mdg1* | *Ele61* | 15 | 6(2) | 6 | 0 |
| *Mdg1* | *Ele125* | 20 | 3(1) | 11 | 2 |
| *Mdg1* | *Ele126* | 18 | 5(2) | 8 | 3 |
| *Mdg1* | *Ele127* | 23 | 5(3) | 14 | 2 |
| *Mdg1* | *Ele137* | 10 | 6(1) | 2 | 1 |
| *Mdg1* | *Ele150* | 7 | 1(0) | 3 | 2 |
| *Mdg1* | *Ele152* | 22 | 2(2) | 10 | 7 |
| *Mdg3* | *Ele13* | 13 | 3(2) | 7 | 2 |
| *Mdg3* | *Ele14* | 11 | 1(1) | 5 | 5 |
| *Mdg3* | *Ele15* | 29 | 3(3) | 16 | 5 |
| *Mdg3* | *Ele16* | 4 | 0 | 3 | 0 |
| *Mdg3* | *Ele17* | 8 | 3(2) | 2 | 2 |
| *Mdg3* | *Ele18* | 6 | 0 | 5 | 1 |
| *Mdg3* | *Ele19* | 8 | 2(1) | 3 | 3 |
| *Mdg3* | *Ele20* | 10 | 1(1) | 6 | 2 |
| *Mdg3* | *Ele21* | 7 | 2(1) | 2 | 2 |
| *Mdg3* | *Ele22* | 10 | 1(0) | 6 | 2 |
| *Mdg3* | *Ele23* | 8 | 2(1) | 6 | 0 |
| *Mdg3* | *Ele24* | 11 | 4(0) | 3 | 4 |
| *Mdg3* | *Ele25* | 11 | 3(2) | 5 | 2 |
| *Mdg3* | *Ele139* | 5 | 1(1) | 4 | 0 |
| *Mdg3* | *Ele140* | 7 | 1(0) | 3 | 3 |
| TOTAL | 52 | 760 | 127(53) | 404 | 144 |

[a] Número de copias activas. Entre paréntesis se indica el número de copias activas que presentan LTRs idénticas.

**Tabla 6**. Media de la identidad entre los elementos activos y entre los inactivos comparados dos a dos, para las familias del grupo *Ty3/gypsy* en *Aegypti*.

| Family[a] | $Id_{act}$[b] | Range[c] | N[d] | $Id_{in}$[e] | Range[c] | N[d] |
|---|---|---|---|---|---|---|
| Ele1 | 99.7 | 99.5-100 | 6 | 92.5 | - | 1 |
| Ele3 | 99.8 | 99.7-100 | 10 | 92.1 | 85.7-98.8 | 25 |
| Ele4 | 99.9 | - | 1 | 98.0 | 97.4-98.6 | 3 |
| Ele5 | 99.3 | - | 1 | 93.9 | 90.9-95.7 | 3 |
| Ele8 | 99.6 | - | 1 | 90.2 | 98.1-98.7 | 64 |
| Ele10 | 99.2 | - | 1 | 94.0 | 91.5-96.0 | 10 |
| Ele11 | 99.4 | - | 1 | 96.1 | 95.7-96.4 | 3 |

[a] Nombres abreviados de las familias.
[b] Media de la identidad de las comparaciones dos a dos de los elementos activos de la familia.
[c] Rango del total de valores de identidad observados.
[d] Número total de comparaciones.
[e] Media de la identidad de las comparaciones dos a dos de los elementos inactivos de la familia (se excluyen Solo-LTRs).

## IV.3. Fijación por menor presión selectiva en regiones heterocromáticas.

Estudios llevados a cabo en los genomas de *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster* y *Tetraodon nigroviridis*, indican una mayor tendencia de los TEs a ser más abundantes en regiones con efectos menos deletéreos y/o donde se espera que la selección natural sea menos eficaz para eliminarlos (Bartolome, Maside & Charlesworth, 2002; Kaminker et al., 2002; Rizzon et al., 2002; Wright, Agrawal & Bureau, 2003; Fischer et al., 2004). En esta misma línea, Holt et al. (2002) estimaron que dentro del componente eucromático del genoma de *A. gambiae*, la densidad de estos elementos repetitivos es mayor cerca de los centrómeros, menor en el medio de brazos cromosómicos, y algo elevada cerca de los telómeros.

La acumulación de TEs en regiones heterocromáticas es un hecho común a varios organismos eucariotas (Bartolomé et al., 2002; Kaminker et al., 2002; Rizzon et al., 2002; Kapitonov & Jurka, 2003; Wright et al., 2003; Fischer et al., 2004). En *D. melanogaster* este hecho ha sido explicado por la supresión de recombinación asociada a dichas regiones. Las hipótesis derivadas de la teoría del ADN egoísta (Doolitle & Sapienza, 1980; Orgel & Crick, 1980) sugieren que una carencia de recombinación permitiría la acumulación de TEs en regiones con baja tasa de recombinación (o con recombinación reducida) debido a que la probabilidad de que los TEs se impliquen en fenómenos de recombinación ectópica que lleven a reordenamientos cromosómicos deletéreos sería mucho menor (Langley et al., 1988; Montgomery et al., 1991). En el genoma de *A. gambiae*, Holt et al. (2002) reportaron

que los TEs constituyen un 60% del componente heterocromático y que mayoritariamente se encuentran altamente fragmentados, y apuntaron, además, que la distribución de los TEs en *A. gambiae* es consecuente con la hipótesis que supone una mayor densidad de estas secuencias en aquellas regiones del genoma donde la tasa de recombinación es menor. Es más, los resultados obtenidos por nuestro grupo de investigación soporta esta idea, al encontrar una significativa sobre-representación de elementos inactivos dentro de los *scaffolds* no mapeados de la cepa PEST de *A. gambiae*, sugiriendo que la heterocromatina es refugio para estas copias degeneradas, debido a una reducción de la acción de la selección natural en contra de los elementos insertados en regiones heterocromáticas (Tubío, Naveira & Costas, 2005).

En *D. melanogaster*, el grupo de Maside (Bartolomé & Maside, 2004; Maside, Assimacopoulos & Charlesworth, 2005) obtuvo resultados que sugieren que los TEs tienden a la fijación en regiones no-recombinantes del genoma, particularmente en el cromosoma *4* (donde los niveles de recombinación son muy bajos). Teniendo en cuenta la estima de que un 60% de la heterocromatina de *A. gambiae* esta conformada por TEs mayoritariamente fragmentados (Holt et al., 2002), es muy posible que las altas tasas de ocupación de los tres *loci* estudiados en el presente trabajo sean debidas a un efecto de este tipo. Siguiendo esta idea, se ha tratado de ubicar los tres *loci* con ayuda del ENSEMBL. Los resultados llevan a pensar que, muy probablemente, el *locus1* y el *locus3* (localizados respectivamente en los cromosomas *3R* y *2L*) se encuentran ubicados en la eucromatina, dado que presentan genes próximos a una distancia de ~25Kb. El *locus2*, sin embargo, podría tener una ubicación heterocromática, teniendo en cuenta que todavía no ha podido ser asociado a un cromosoma. El número de *loci* analizados en esta investigación no es suficiente para poder hacer conclusiones a gran escala. Sin embargo, la identificación de dos inserciones fijadas que probablemente están ubicadas en la eucromatina, lleva a pensar que la posibilidad de encontrar inserciones fijadas en la heterocromatina, donde se encuentran la mayoría de elementos degenerados del genoma de *A. gambiae*, sea mayor.

### IV.4. Fijación por asociación a inversiones con valor adaptativo.

Los resultados expuestos sugieren que la diferente tasa de renovación observada para *A. gambiae* y *D. melanogaster* no se debe a diferencias en la densidad génica de

ambos genomas. Sin embargo, consecuente con la idea que supone una relajación de la acción purificadora de la selección natural para explicar la menor tasa de renovación global de retrotransposones en el genoma de *A. gambiae*, y teniendo en cuenta la importancia de las inversiones en la evolución de las especies del Complejo de *A. gambiae* (Ayala & Coluzzi, 2005), puede proponerse como factor adicional el ligamiento a reordenamientos cromosómicos con valor adaptativo.

Las inversiones, así como otros reordenamientos cromosómicos, pueden jugar un importante papel en la diferenciación ecológica entre poblaciones parapátricas y simpátricas de una especie y, por tanto, en la especiación, como consecuencia de la supresión de recombinación entre reordenamientos alternativos y la estabilización de combinaciones alélicas con valor adaptativo (Powell et al., 1999). Coluzzi propuso un modelo de recombinación suprimida para explicar la especiación dentro del Complejo de *A. gambiae* (revisión por Ayala & Coluzzi, 2005). Según este modelo, inicialmente una población ancestral cromosómicamente monomórfica se expandiría en número, tras colonizar un ambiente con condiciones favorables. Esta población inicial encontraría nuevas condiciones ambientales, bien en los márgenes del rango de distribución de dicha especie, o bien dentro del propio rango de distribución como consecuencia de oscilaciones ecológicas estacionales o climáticas. Las mutaciones que suponen una mejor adaptación a estas condiciones ecológicas marginales serán favorecidas frente a otras variantes que supongan peor adaptación, produciendo nuevos ecotipos[h]. La oscilación de estas condiciones ambientales marginales tendrá el mismo efecto sobre el ecotipo adaptado a las mismas, resultando en cambios oscilantes en el tamaño de dicha subpoblación. La reproducción entre los individuos de esta subpoblación y los individuos propios de la población central permitirá la difusión de alelos adaptados a las condiciones marginales, sin embargo no difundirían en la población central aquellos alelos que hayan quedado "encerrados" dentro de un reordenamiento cromosómico surgido en la subpoblación, como consecuencia de la supresión de la recombinación entre heterocariotipos. Así, el cruzamiento entre individuos de la población central y la población marginal en aquellas áreas parapátricas homogeneizará genéticamente ambas poblaciones, exceptuando los alelos protegidos por los reordenamientos cromosómicos, donde se podrán acumular nuevos alelos o combinaciones de alelos con valor adaptativo, incluyendo alelos que

[h] grupos intraespecíficos que presentan caracteres específicos que resultaron como consecuencia de la presión selectiva en unas condiciones ambientales locales.

promoverán el posterior aislamiento reproductivo. Este aislamiento reproductivo surgirá gradualmente, dando lugar primero a una especiación incipiente y, eventualmente, especies claramente definidas.

La preponderancia de TEs degenerados-inactivos en el genoma de *A. gambiae* y la posible fijación (o, al menos, alta tasa de ocupación) de muchas de estas inserciones sugerida en el presente trabajo, podría ser consecuencia del *hitchhiking* de inserciones asociadas a estas inversiones con valor adaptativo tan importantes durante el proceso de especiación en el Complejo de *A. gambiae*. La idea general de esta propuesta es la misma que la sugerida por Sniegowski & Charlesworth (1994), cuando encontraron altas tasas de ocupación de TEs localizados dentro de las inversiones In(3R)Mo e In(3R)K de *Drosophila* y apuntaron al *hitchhiking* como responsable del aumento de dichas frecuencias de ocupación. Por un lado, la supresión (o reducción) de la recombinación en las regiones asociadas a las inversiones entre los heterocariotipos reducirá significativamente los efectos deletéreos por recombinación ectópica promovidos por los TEs, reduciendo los efectos de la selección purificadora. Por otro lado, el valor adaptativo de determinados *loci*, o combinaciones de *loci*, asociados a estas inversiones conferirán una ventaja selectiva a los individuos que presenten el reordenamiento, de manera que la frecuencia del mismo se vería incrementado en la subpoblación, estabilizándose la combinación de los *loci* que suponen la ventaja adaptativa, y también aquellos *loci* localizados dentro de la inversión que, por efecto *hitchhiking*, se verán conducidos hacia la fijación en la subpoblación.

### IV.5. Fijación por reducción del tamaño efectivo de población.

La hipótesis de la asociación a inversiones con valor adaptativo para explicar la alta tasa de ocupación de TEs en *A. gambiae* es consistente con el modelo de especiación propuesto por Coluzzi para el Complejo de *A. gambiae*. Sin embargo, desde luego no parece que sea el único factor, aunque probablemente sí uno de los dos más importantes. Como segundo factor principal proponemos la reducción en el tamaño efectivo de población. Los datos generados tras el estudio de la evolución del complemento TE de la especie *Homo sapiens* han desvelado que $N_e$ es la clave para entender el patrón exhibido por los elementos móviles en nuestra especie. Los TEs constituyen ~50% del tamaño total

del genoma humano que, mayoritariamente, constituyen inserciones inactivas y muy antiguas (Lower, Lower & Kurth, 1996; Medstrand & Mager, 1998; Tristem, 2000) que se encuentran fijadas, viéndose reducidos sus potenciales efectos deletéreos. El reducido tamaño efectivo de población que ha caracterizado la historia evolutiva de nuestra especie y los consecuentes cuellos de botella generados (Takahata, 1993) explican que la selección no haya eliminado este *pool* de inserciones antiguas que forman parte de nuestro genoma.

Por extensión de este modelo, la estructuración que caracteriza al Complejo de especies de *A. gambiae* y, en concreto, a la especie *A. gambiae*, puede estar detrás del patrón global que los TEs exhiben en el genoma del mosquito. Es más, el papel que esta fuerza evolutiva ha podido jugar en la fijación de TEs cobra mayor importancia si se atiende a que los efectos demográficos se reflejan en la globalidad del genoma, mientras que los efectos del *hitchhiking* en inversiones adaptativas están restringidos a las regiones propias o próximas a los reordenamientos cromosómicos (Depaulis et al., 2000; Andolfatto, 2001). Es más, Stump et al. (2007) estimaron que la recombinación entre los heterocariotipos para la inversión adaptativa *2La* de *A. gambiae* tiene lugar a una tasa de 0.5 cM/Mb, proponiendo que, por extensión a otras inversiones que tienen o hayan tenido lugar en las especies del Complejo, el modelo de supresión de recombinación propuesto por Coluzzi no sería suficiente para explicar la especiación dentro del complejo de *A. gambiae*. Una tasa de recombinación del orden estimado por Stump et al. (2007) podría reducir los efectos deletéreos de los TEs derivados de la recombinación ectópica, pero nunca suprimirlos en la misma cuantía que supondría la supresión total de recombinación. Por tanto, todo parece indicar que otras fuerzas han jugado un papel importante en la configuración del patrón global mostrado por los TEs en el genoma de *A. gambiae*, y una de estas probablemente haya sido la reducción del tamaño eficaz de las poblaciones a lo largo de la historia evolutiva de la especie.

La estructura poblacional de la especie *Anopheles gambiae* es muy compleja. El nivel taxonómico más alto del sistema es *Anopheles gambiae sensu lato*, que comprende al menos siete especies morfológicamente indistinguibles y entre las cuales existe cierto grado de flujo génico. Una de estas especies es *Anopheles gambiae sensu estricto*, que presenta hasta cinco formas cromosómicas con diferentes niveles de tolerancia a distintas condiciones ecológicas. También existe un proceso de especiación incipiente dentro de la especie, distinguiéndose las formas moleculares *M* y *S* de *A. gambiae*, que presentan

restricción de flujo génico entre ambas. Además, los cambios estacionales anuales propios del clima, los largos períodos de sequía y la presencia de infranqueables barreras físicas, como la representada por el impenetrable Valle del Rift, son factores que determinan la estructuración poblacional de esta especie en África (Krzywinski & Besansky, 2003). A pesar de todo, los datos demográficos actuales descartan la posibilidad de cuellos de botella en la historia evolutiva reciente de la especie, excepto para el Este de Kenia, como consecuencia del Valle del Rift (Lehmann et al., 2000), ya que se ha estimado que el tamaño efectivo de población de *A. gambiae* a largo plazo sería, probablemente, del orden de los cientos de miles de individuos (Lehmann et al., 1998). El origen de las primeras divergencias dentro del Complejo se remontan a ~5000 años. Ayala & Coluzzi (2005) proponen que *A. arabiensis*, presumiblemente la especie a partir de la cual surgió el Complejo de *A. gambiae*, descendería de una especie zoofílica y exofílica ancestral del género *Pyretophorus* procedente de la Península arábiga. Nada sabemos sobre la estructura genética de las poblaciones de estas especies ancestrales. Sin embargo, la baja tasa de renovación observada para los retrotransposones de *A. gambiae*, así como la probable fijación de la mayoría de estas inserciones degeneradas y antiguas, es compatible con los efectos de la deriva genética sobre estas poblaciones ancestrales.

# V. Resumen y Conclusiones

## RESUMEN

Los elementos genéticos transponibles (abreviadamente, TEs) son un componente importante de muchos genomas eucariotas. Originariamente fueron considerados secuencias de ADN con un comportamiento estrictamente parasítico, con efectos deletéreos sobre los genomas que los hospedan como consecuencia de su actividad transposicional (Charlesworth, Sniegowsky & Stephan, 1994). Sin embargo, a lo largo de la última década se han ido recopilando gran cantidad de evidencias a nivel molecular que han terminado por demostrar, de manera inequívoca, la inmensidad de caminos a través de los cuales los TEs pueden impactar en la evolución de los genomas y, por tanto, de las especies (Kidwell & Lisch, 2001). El creciente interés por conocer con mayor exactitud este impacto estructural y funcional sobre los genomas ha cobrado mayor repercusión con la accesibilidad a los primeros genomas secuenciados, llevando a algunos investigadores del campo de la evolución molecular a analizar este complemento transposicional, desde los TEs más recientes y potencialmente activos hasta los restos de aquellos elementos transponibles más deteriorados y antiguos.

El objetivo inicial de esta tesis era obtener unos primeros datos acerca de la dinámica de los retrotransposones del grupo *Ty3/gypsy* en el genoma del mosquito *Anopheles gambiae*, y compararla con el patrón exhibido por otros genomas, especialmente por el de *Drosophila melanogaster*. Para obtener estas primeras comparaciones, decidimos centrarnos en el linaje de *Mdg1* de dicho grupo de retroelementos. Se eligió este linaje por dos motivos: el primero es que el grupo *Ty3/gypsy* resultara ser el grupo de retrotransposones más abundante en el genoma de *D. melanogaster* (Kaminker et al., 2002); el segundo motivo era que ya disponíamos de un extenso trabajo sobre el linaje de *Mdg1* en *D. melanogaster*, desarrollado en años anteriores por nuestro grupo de investigación (Costas, Valadé & Naveira, 2001), lo que sin duda facilitaría mucho el desarrollo de la investigación.

Esta primera investigación, sobre el linaje de *Mdg1* en el genoma de *A. gambiae,* reveló un patrón claramente diferente al conocido para *D. melanogaster*. La proporción de inserciones fragmentadas y de LTRs solitarias era significativamente mayor en el genoma del mosquito frente al de *D. melanogaster* (Tubío, Costas & Naveira, 2004). Tras extender

el estudio al resto de linajes del Grupo *Ty3/gypsy*, pudimos confirmar que el patrón observado para los elementos del linaje *Mdg1* podía aplicarse a la globalidad de los elementos del grupo (linajes *Gypsy*, *Mdg3*, *CsRn1*, *Osvaldo* y *Mag*), es decir, una clara preponderancia de copias inactivas, mayoritariamente fragmentadas, sobre las activas. Además, estas copias fragmentadas presentaban una divergencia entre sí significativamente mayor a la que presentaban las copias activas entre sí, algo que rompía con la homogeneidad exhibida por los retrotransposones de *D. melanogaster*. Estos resultados nos llevaron a reportar una menor tasa de renovación de retrotransposones en el genoma de *A. gambiae* con respecto al de *D. melanogaster* (Tubío, Naveira & Costas, 2005).

Para explicar esta menor tasa de renovación de retrotransposones en el genoma de *A. gambiae*, nuestra hipótesis de partida era que la mayoría de estas inserciones serían componentes antiguos del genoma de *Anopheles* y que, por tanto, su persistencia en este genoma se debería a que dichas inserciones se encontrarían fijadas. Para testar esta hipótesis, llevamos nuestro trabajo al *Centre for Tropical Disease Research and Training* (University of Notre Dame), donde se nos proporcionaría el material y las muestras necesarias para estimar la tasa de ocupación de algunas inserciones de retrotransposones en poblaciones naturales de *A. gambiae*.

Dado que el objetivo prioritario del estudio era encontrar inserciones fijadas en *A. gambiae*, llevamos a cabo una selección de *loci* candidatos, tomando como principales criterios dos: (1) deberían ser loci antiguos, y (2) deberían ser de tamaño pequeño. Atendiendo a estas dos premisas, se filtraron 642 TEs de una base de datos previamente elaborada por nuestro grupo de investigación (Tubío, Naveira & Costas, 2005), de la que se seleccionaron los 10 *loci* que mejor se ajustaban a los criterios de partida. Finalmente, tras varias pruebas de optimización de las PCRs sobre estos 10 *loci* candidatos, se escogieron aquellos 3 *loci* con mejores resultados en la amplificación. El genotipado de los tres *loci* estudiados permitió estimar una tasa de ocupación del 100%, habiendo sido detectados en todos los individuos de las poblaciones y subpoblaciones analizadas que cubrían toda la estructura de las poblaciones de *A. gambiae*. Es más, esta observación fue verificada mediante la secuenciación, confirmando que la secuencia de los límites (5´ y 3´) de cada uno de los *loci* resultó ser más de un 99% idéntico a la secuencia publicada del genoma de la cepa PEST de *A. gambiae* (Holt et al., 2002). Además, la comparación de la secuencia total obtenida para los tres *loci* en las poblaciones y subpoblaciones estudiadas

es consistente con la idea de que éstos se encuentran fijados en el genoma del mosquito *A. gambiae*. Es más, la secuenciación de 2 de estos *loci* en la especie *A. arabiensis* (que representa la especie basal dentro del Complejo de *Anopheles gambiae*) reveló que ambos también se encuentran presentes en el genoma de esta especie.

La fijación de estas tres inserciones antiguas en el genoma de *A. gambiae* era consistente con nuestra hipótesis de partida (Tubío, Costas & Naveira, 2004) que suponía altas tasas de ocupación en aquellas inserciones más antiguas que no han sido eliminadas por la acción de la selección. Estos resultados, compatibles con las suposiciones que establecen los modelos derivados de la teoría sobre la evolución del ADN parasítico, podrían ser interpretados principalmente como (1) una relajación de la acción purificadora de la selección natural, como consecuencia de la menor densidad génica del genoma de *A. gambiae* con respecto al genoma de *D. melanogaster*, este último mucho más compacto; (2) una relajación de la selección natural por la asociación de inserciones a reordenamientos cromosómicos con recombinación reducida, como son las inversiones paracéntricas tan comunes en el genoma de *Anopheles*; (3) un tamaño efectivo de población reducido durante la historia evolutiva de la especie, que haya supuesto la acción de la deriva genética.

La reciente secuenciación del genoma del mosquito *Aedes aegypti*, supuso una extraordinaria oportunidad para testar la primera de las suposiciones indicadas. Este genoma tiene una densidad génica menor que el de *A. gambiae*, como consecuencia del elevado tamaño de su genoma (1.3Gb frente a las 278Mb de A. gambiae). Las estimas realizadas sobre la tasa de renovación de retrotransposones en *Aedes* resultaron no diferir significativamente de las tasas reportadas para *Anopheles*. Por extensión, estos resultados nos llevaron a rechazar la hipótesis de la diferente densidad génica para explicar las diferencias entre *A. gambiae* y *D. melanogaster*.

Tras descartar esta primera idea nos decantamos por proponer las otras dos como más probables. Por un lado, la asociación de elementos móviles con inversiones adaptativas, podría arrastrar a muchas inserciones hacia la fijación por efecto *hitchhiking*, teniendo en cuenta que la especiación en el complejo de *A. gambiae* se ha venido explicando principalmente por un modelo de recombinación suprimida (Ayala & Coluzzi, 2005). Por otro lado, el importante nivel de estructuración que caracteriza al Complejo de *A. gambiae* y, en general, aplicable a otras especies del Género, así como las reducciones

del tamaño de población observadas durante los cambios estacionales, han podido implicar la acción de la deriva genética sobre sus genomas. Lamentablemente, los datos sobre la historia evolutiva de las poblaciones de *Anopheles*, así como la configuración de los reordenamientos de las especies ancestrales del Complejo, no permiten estimar la implicación relativa de ambas fuerzas.

## CONCLUSIONES

**1.** Existen al menos 73 familias bien representadas de retrotransposones con LTRs del grupo *Ty3/gypsy* en *A. gambiae*, repartidas en cinco de los seis linajes conocidos en insectos: *Gypsy*, *Mdg1*, *Mdg3*, *CsRn1* y *Mag*. El linaje de *Osvaldo* parece no estar representado, ni siquiera por inserciones degeneradas. A pesar de todo, existe una importante abundancia y diversidad en comparación con el genoma de *D. melanogaster*.

**2.** Esta importante diversidad también se ve reflejada a nivel estructural.

**3.** En *Ae. aegypti* existen al menos 184 familias bien representadas de retrotransposones del grupo *Ty3/gypsy*. La abundancia de familias en comparación con *A. gambiae* está claramente relacionada con la diferencia en el tamaño de ambos genomas. Todos los linajes del grupo *Ty3/gypsy* de insectos están bien representados en *Aedes*: *Gypsy*, *Mdg1*, *Mdg3*, *CsRn1*, *Mag* y *Osvaldo*.

**4.** No se han detectado eventos de transferencia horizontal en el genoma de *A. gambiae*. Sin embargo, se ha confirmado que la evolución en mosaico, debida probablemente a recombinación, juega un papel destacable en la evolución de los retrotransposones del grupo *Ty3/gypsy* de *A. gambiae*.

**5.** El genoma de *A. gambiae* muestra, en comparación con la renovación observada en *D. melanogaster*, una baja tasa de renovación de retrotransposones, caracterizada por una clara preponderancia de inserciones inactivas (incluidas LTRs solitarias) divergentes, con respecto a inserciones activas más homogéneas.

**6.** La formación de LTRs solitarias, como consecuencia de la recombinación de ambas LTRs de un mismo retrotransposón, podría estar jugando un importante papel en la inactivación de los retrotransposones con LTRs de *A. gambiae*, ya que parece que existe una relación directa entre el número de LTRs solitarias y la longitud de dichas LTRs. Esto podría sugerir que la persistencia en el genoma de *A. gambiae* de aquellas familias con mayor longitud de LTR (como el linaje *Osvaldo*) se vería amenazada por la mayor probabilidad de recombinación.

**7.** La tasa de renovación de retrotransposones en el genoma de *Ae. aegypti* no parece diferir de la observada en el de *A. gambiae*, puesto que no existen diferencias significativas en cuanto al número de inserciones inactivas *versus* activas entre ambos genomas. Además, la divergencia entre copias inactivas del genoma de *Ae. aegypti* es significativamente mayor que la divergencia entre copias activas. Estos datos de divergencia tampoco parecen diferir de los observados en el genoma de *A. gambiae*.

**8.** La preponderancia de inserciones inactivas y divergentes de los retrotransposones del genoma de *A. gambiae* nos lleva a pensar que la mayoría de estas sean antiguas y que por tanto, posiblemente, muchas de ellas se encuentren fijadas en el genoma del mosquito. El estudio de la tasa de ocupación de tres inserciones antiguas seleccionadas en el genoma de *A. gambiae*, reveló que se encontraban fijadas, lo que refuerza la hipótesis de la fijación global de muchas de las inserciones inactivas más divergentes. Además, al menos dos de las inserciones estudiadas se encuentran presentes en *Anopheles arabiensis*, una de ellas posiblemente fijada en *A. arabiensis*.

**9.** La comparación del grado de actividad mostrado por los genomas de *A. gambiae* y *Ae. aegypti*, permite concluir que la diferente densidad génica de ambos genomas no parece jugar un papel importante en la acumulación de inserciones inactivas. Por extensión, la diferente densidad génica observada entre *A. gambiae* y *D. melanogaster* podría no ser responsable del diferente patrón de actividad observado para los retrotransposones de ambos genomas.

**10.** Como camino a seguir en futuras investigaciones de cara a explicar la baja tasa de renovación de retrotransposones de *A. gambiae* y la presumible alta tasa de ocupación de muchas de sus inserciones, apuntamos como causas más probables dos: (1) el *hitchhiking* de inserciones asociadas a aquellas inversiones que encierran combinaciones alélicas con

valor adaptativo; y (2) la estructuración genética que caracterizaría a las poblaciones ancestrales de la línea evolutiva que dio lugar al Complejo de *A. gambiae*.

# VI. Bibliografía

Agrawal, A., Q. M. Eastman, et al. (1998). "Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system." Nature 394(6695): 744-51.

Andolfatto, P. (2001). "Adaptive hitchhiking effects on genome variability." Curr Opin Genet Dev 11(6): 635-41.

Aquadro, C. F., A. L. Weaver, et al. (1991). "Molecular evolution of inversions in Drosophila pseudoobscura: the amylase gene region." Proc Natl Acad Sci U S A 88(1): 305-9.

Ayala, F. J. and M. Coluzzi (2005). "Chromosome speciation: humans, Drosophila, and mosquitoes." Proc Natl Acad Sci U S A 102 Suppl 1: 6535-42.

Bartolome, C., X. Maside, et al. (2002). "On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster." Mol Biol Evol 19(6): 926-37.

Bergman, C. M. and D. Bensasson (2007). "Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in Drosophila melanogaster." Proc Natl Acad Sci U S A 104(27): 11340-5.

Besansky, N. J., J. R. Powell, et al. (1994). "Molecular phylogeny of the Anopheles gambiae complex suggests genetic introgression between principal malaria vectors." Proc Natl Acad Sci U S A 91(15): 6885-8.

Besansky, N. J., T. Lehmann, et al. (1997). "Patterns of mitochondrial variation within and between African malaria vectors, Anopheles gambiae and An. arabiensis, suggest extensive gene flow." Genetics 147(4): 1817-28.

Biedler, J. and Z. Tu (2003). "Non-LTR retrotransposons in the African malaria mosquito, Anopheles gambiae: unprecedented diversity and evidence of recent activity." Mol Biol Evol 20(11): 1811-25.

Biessmann, H. and J. M. Mason (1997). "Telomere maintenance without telomerase." Chromosoma 106(2): 63-9.

Biessmann, H., M. F. Walter, et al. (1999). "Moose, a new family of LTR-retrotransposons in the mosquito Anopheles gambiae." Insect Mol Biol 8(2): 201-12.

Black, W. C. t. and G. C. Lanzaro (2001). "Distribution of genetic variation among chromosomal forms of Anopheles gambiae s.s: introgressive hybridization, adaptive inversions, or recent reproductive isolation?" Insect Mol Biol 10(1): 3-7.

Boeke. J, D., H. Eickbush. T, et al. (1998). Virus taxonomy: ICTV VIIth Report. New York, Springer-Verlag. Ed.

Bourc'his, D. and T. H. Bestor (2004). "Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L." Nature 431(7004): 96-9.

Bowen, N. J. and J. F. McDonald (2001). "Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside." Genome Res 11(9): 1527-40.

Breman, J. G., A. Egan, et al. (2001). "The intolerable burden of malaria: a new look at the numbers." Am J Trop Med Hyg 64(1-2 Suppl): iv-vii.

Bryan, J. H., V. Petrarca, et al. (1987). "Adult behaviour of members of the Anopheles gambiae complex in the Gambia with special reference to An. melas and its chromosomal variants." Parassitologia 29(2-3): 221-49.

Buhler, M., A. Verdel, et al. (2006). "Tethering RITS to a nascent transcript initiates RNAi- and heterochromatin-dependent gene silencing." Cell 125(5): 873-86.

Caccone, A., B. A. Garcia, et al. (1996). "Evolution of the mitochondrial DNA control region in the Anopheles gambiae complex." Insect Mol Biol 5(1): 51-9.

Caccone, A., G. S. Min, et al. (1998). "Multiple origins of cytologically identical chromosome inversions in the Anopheles gambiae complex." Genetics 150(2): 807-14.

Caceres, M., J. M. Ranz, et al. (1999). "Generation of a widespread Drosophila inversion by a transposable element." Science 285(5426): 415-8.

Calvi, B. R., T. J. Hong, et al. (1991). "Evidence for a common evolutionary origin of inverted repeat transposons in Drosophila and plants: hobo, Activator, and Tam3." Cell 66(3): 465-71.

Cao, X. and S. E. Jacobsen (2002). "Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing." Curr Biol 12(13): 1138-44.

Cappello, J., K. Handelsman, et al. (1985). "Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence." Cell 43(1): 105-15.

Capy, P., T. Langin, et al. (1997). "Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor?" Genetica 100(1-3): 63-72.

Capy, P. (1998). "Evolutionary biology. A plastic genome." Nature 396(6711): 522-3.

Capy, P., C. Bazin, et al. (1998). Dynamics and evolution of transposable elements. Austin.

Carnahan, J., L. Zheng, et al. (2002). "Genetic differentiation of Anopheles gambiae s.s. populations in Mali, West Africa, using microsatellite loci." J Hered 93(4): 249-53.

Civardi, L., Y. Xia, et al. (1994). "The relationship between genetic and physical distances in the cloned a1-sh2 interval of the Zea mays L. genome." Proc Natl Acad Sci U S A 91(17): 8268-72.

Coghlan, A., E. E. Eichler, et al. (2005). "Chromosome evolution in eukaryotes: a multi-kingdom perspective." Trends Genet 21(12): 673-82.

Colot, H. V., J. C. Hall, et al. (1988). "Interspecific comparison of the period gene of Drosophila reveals large blocks of non-conserved coding DNA." EMBO J 7(12): 3929-37.

Coluzzi, M., A. Sabatini, et al. (1979). "Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex." Trans R Soc Trop Med Hyg 73(5): 483-97.

Coluzzi, M., V. Petrarca, et al. (1985). "Chromosomal inversion intergradation and incipient speciation in Anopheles gambiae." Boll. Zool. 52: 45-63.

Coluzzi, M., A. Sabatini, et al. (2002). "A polytene chromosome analysis of the Anopheles gambiae species complex." Science 298(5597): 1415-8.

Collins, F. H. and S. M. Paskewitz (1996). "A review of the use of ribosomal DNA (rDNA) to differentiate among cryptic Anopheles species." Insect Mol Biol 5(1): 1-9.

Cook, J. M., J. Martin, et al. (2000). "Systematic screening of Anopheles mosquito genomes yields evidence for a major clade of Pao-like retrotransposons." Insect Mol Biol 9(1): 109-17.

Copeland, C. S., V. H. Mann, et al. (2005). "The Sinbad retrotransposon from the genome of the human blood fluke, Schistosoma mansoni, and the distribution of related Pao-like elements." BMC Evol Biol 5(1): 20.

Corces, V. G. and P. K. Geyer (1991). "Interactions of retrotransposons with the host genome: the case of the gypsy element of Drosophila." Trends Genet 7(3): 86-90.

Cornel, A. J. and F. H. Collins (2000). "Maintenance of chromosome arm integrity between two Anopheles mosquito subgenera." J Hered 91(5): 364-70.

Costas, J., E. Valade, et al. (2001). "Amplification and phylogenetic relationships of a subfamily of blood, a retrotransposable element of Drosophila." J Mol Evol 52(4): 342-50.

Courage, U., H. P. Doring, et al. (1984). "Transposable elements Ac and Ds at the shrunken, waxy, and alcohol dehydrogenase 1 loci in Zea mays L." Cold Spring Harb Symp Quant Biol 49: 329-38.

Cresse, A. D., S. H. Hulbert, et al. (1995). "Mu1-related transposable elements of maize preferentially insert into low copy number DNA." Genetics 140(1): 315-24.

Currie, D. B., T. F. Mackay, et al. (1998). "Pervasive effects of P element mutagenesis on body size in Drosophila melanogaster." Genet Res 72(1): 19-24.

Charlesworth, B. and C. H. Langley (1986). "The evolution of self-regulated transposition of transposable elements." Genetics 112(2): 359-83.

Charlesworth, B. and C. H. Langley (1989). "The population genetics of Drosophila transposable elements." Annu Rev Genet 23: 251-87.

Charlesworth, B., P. Sniegowski, et al. (1994). "The evolutionary dynamics of repetitive DNA in eukaryotes." Nature 371(6494): 215-20.

Charlesworth, B., C. H. Langley, et al. (1997). "Transposable element distributions in Drosophila." Genetics 147(4): 1993-5.

Daboussi, M. J. and P. Capy (2003). "Transposable elements in filamentous fungi." Annu Rev Microbiol 57: 275-99.

Daniels, S. B. and A. Chovnick (1993). "P element transposition in Drosophila melanogaster: an analysis of sister-chromatid pairs and the formation of intragenic secondary insertions during meiosis." Genetics 133(3): 623-36.

Davidson, G. (1964). "Anopheles Gambiae, a Complex of Species." Bull World Health Organ 31: 625-34.

de Chastonay, Y., H. Felder, et al. (1992). "Nucleotide sequence of PAT, a retroid element with unusual DR organization, isolated from Panagrellus redivivus." DNA Seq 3(4): 251-5.

Deininger, P. L. and M. A. Batzer (1999). "Alu repeats and human disease." Mol Genet Metab 67(3): 183-93.

della Torre, A., L. Merzagora, et al. (1997). "Selective introgression of paracentric inversions between two sibling species of the Anopheles gambiae complex." Genetics 146(1): 239-44.

della Torre, A., C. Fanello, et al. (2001). "Molecular evidence of incipient speciation within Anopheles gambiae s.s. in West Africa." Insect Mol Biol 10(1): 9-18.

della Torre, A., C. Costantini, et al. (2002). "Speciation within Anopheles gambiae--the glass is half full." Science 298(5591): 115-7.

della Torre, A., Z. Tu, et al. (2005). "On the distribution and genetic differentiation of Anopheles gambiae s.s. molecular forms." Insect Biochem Mol Biol 35(7): 755-69.

DeMarco, R., T. M. Venancio, et al. (2006). "SmTRC1, a novel Schistosoma mansoni DNA transposon, discloses new families of animal and fungi transposons belonging to the CACTA superfamily." BMC Evol Biol 6: 89.

Depaulis, F., L. Brazier, et al. (2000). "Selective sweep near the In(2L)t inversion breakpoint in an African population of Drosophila melanogaster." Genet Res 76(2): 149-58.

Dewannieux, M., C. Esnault, et al. (2003). "LINE-mediated retrotransposition of marked Alu sequences." Nat Genet 35(1): 41-8.

Dimitri, P. (1997). "Constitutive heterochromatin and transposable elements in Drosophila melanogaster." Genetica 100(1-3): 85-93.

Dimitri, P. and N. Junakovic (1999). "Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin." Trends Genet 15(4): 123-4.

Donnelly, M. J. and H. Townson (2000). "Evidence for extensive genetic differentiation among populations of the malaria vector Anopheles arabiensis in Eastern Africa." Insect Mol Biol 9(4): 357-67.

Donnelly, M. J., M. C. Licht, et al. (2001). "Evidence for recent population expansion in the evolutionary history of the malaria vectors Anopheles arabiensis and Anopheles gambiae." Mol Biol Evol 18(7): 1353-64.

Doolittle, W. F. and C. Sapienza (1980). "Selfish genes, the phenotype paradigm and genome evolution." Nature 284(5757): 601-3.

Duncan, L., K. Bouckaert, et al. (2002). "kangaroo, a mobile element from Volvox carteri, is a member of a newly recognized third class of retrotransposons." Genetics 162(4): 1617-30.

Dupressoir, A. and T. Heidmann (1996). "Germ line-specific expression of intracisternal A-particle retrotransposons in transgenic mice." Mol Cell Biol 16(8): 4495-503.

Duval-Valentin, G., B. Marty-Cointin, et al. (2004). "Requirement of IS911 replication before integration defines a new bacterial transposition pathway." EMBO J 23(19): 3897-906.

Earnshaw, W. C., K. F. Sullivan, et al. (1987). "Molecular cloning of cDNA for CENP-B, the major human centromere autoantigen." J Cell Biol 104(4): 817-29.

Eickbush, T. H. (1992). "Transposing without ends: the non-LTR retrotransposable elements." New Biol 4(5): 430-40.

Evgen'ev, M. B., H. Zelentsova, et al. (1997). "Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in Drosophila virilis." Proc Natl Acad Sci U S A 94(1): 196-201.

Evgen'ev, M. B. and I. R. Arkhipova (2005). "Penelope-like elements--a new class of retroelements: distribution, function and possible evolutionary significance." Cytogenet Genome Res 110(1-4): 510-21.

Fanti, L., D. R. Dorer, et al. (1998). "Heterochromatin protein 1 binds transgene arrays." Chromosoma 107(5): 286-92.

Favia, G., A. della Torre, et al. (1997). "Molecular identification of sympatric chromosomal forms of Anopheles gambiae and further evidence of their reproductive isolation." Insect Mol Biol 6(4): 377-83.

Favia, G., A. Lanfrancotti, et al. (2001). "Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of Anopheles gambiae s.s." Insect Mol Biol 10(1): 19-23.

Fedoroff, N., M. Schlappi, et al. (1995). "Epigenetic regulation of the maize Spm transposon." Bioessays 17(4): 291-7.

Feschotte, C. (2004). "Merlin, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences." Mol Biol Evol 21(9): 1769-80.

Feschotte, C. and E. J. Pritham (2005). "Non-mammalian c-integrases are encoded by giant transposable elements." Trends Genet 21(10): 551-2.

Finnegan, D. J. (1989). "Eukaryotic transposable elements and genome evolution." Trends Genet 5(4): 103-7.

Fischer, C., L. Bouneau, et al. (2004). "Global heterochromatic colocalization of transposable elements with minisatellites in the compact genome of the pufferfish Tetraodon nigroviridis." Gene 336(2): 175-83.

Fitzpatrick, T. B. (1986). "Ultraviolet-induced pigmentary changes: benefits and hazards." Curr Probl Dermatol 15: 25-38.

Fontdevila, A. and A. Moya (2003). Evolución. Origen adaptación y divergencia de las especies, Ed. Sintesis.

Garcia, B. A., A. Caccone, et al. (1996). "Inversion monophyly in African anopheline malaria vectors." Genetics 143(3): 1313-20.

Gendrel, A. V., Z. Lippman, et al. (2002). "Dependence of heterochromatic histone H3 methylation patterns on the Arabidopsis gene DDM1." Science 297(5588): 1871-3.

Gentile, G., M. Slotman, et al. (2001). "Attempts to molecularly distinguish cryptic taxa in Anopheles gambiae s.s." Insect Mol Biol 10(1): 25-32.

Gillies, N. E. (1961). "The use of auxotrophic mutants to study restoration in Escherichia coli B after ultra-violet-irradiation." Int J Radiat Biol 3: 379-87.

Gillies, C. (1987). "Infant colic: is there anything new?" J Pediatr Health Care 1(6): 305-12.

Goodwin, T. J. and R. T. Poulter (2001). "The DIRS1 group of retrotransposons." Mol Biol Evol 18(11): 2067-82.

Goodwin, T. J., M. I. Butler, et al. (2003). "Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi." Microbiology 149(Pt 11): 3099-109.

Goodwin, T. J. and R. T. Poulter (2004). "A new group of tyrosine recombinase-encoding retrotransposons." Mol Biol Evol 21(4): 746-59.

Gowher, H., O. Leismann, et al. (2000). "DNA of Drosophila melanogaster contains 5-methylcytosine." EMBO J 19(24): 6918-23.

Greene, B., R. Walko, et al. (1994). "Mutator insertions in an intron of the maize knotted1 gene result in dominant suppressible mutations." Genetics 138(4): 1275-85.

Grossman, G. L., A. J. Cornel, et al. (1999). "Tsessebe, Topi and Tiang: three distinct Tc1-like transposable elements in the malaria vector, Anopheles gambiae." Genetica 105(1): 69-80.

Hammer, S. E., S. Strehl, et al. (2005). "Homologs of Drosophila P transposons were mobile in zebrafish but have been domesticated in a common ancestor of chicken and human." Mol Biol Evol 22(4): 833-44.

Hehl, R., W. K. Nacken, et al. (1991). "Structural analysis of Tam3, a transposable element from Antirrhinum majus, reveals homologies to the Ac element from maize." Plant Mol Biol 16(2): 369-71.

Hickey, D. A. (1982). "Selfish DNA: a sexually-transmitted nuclear parasite." Genetics 101(3-4): 519-31.

Higashiyama, T., Y. Noutoshi, et al. (1997). "Zepp, a LINE-like retrotransposon accumulated in the Chlorella telomeric region." EMBO J 16(12): 3715-23.

Hill, S. R., S. S. Leung, et al. (2001). "Ikirara insertions reveal five new Anopheles gambiae transposable elements in islands of repetitious sequence." J Mol Evol 52(3): 215-31.

Hiom, K., M. Melek, et al. (1998). "DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations." Cell 94(4): 463-70.

Holmes, S. E., B. A. Dombroski, et al. (1994). "A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion." Nat Genet 7(2): 143-8.

Holt, R. A., G. M. Subramanian, et al. (2002). "The genome sequence of the malaria mosquito Anopheles gambiae." Science 298(5591): 129-49.

Huang, J., T. Fan, et al. (2004). "Lsh, an epigenetic guardian of repetitive elements." Nucleic Acids Res 32(17): 5019-28.

Hunt, R. H., M. Coetzee, et al. (1998). "The Anopheles gambiae complex: a new species from Ethiopia." Trans R Soc Trop Med Hyg 92(2): 231-5.

Irvine, D. V., M. Zaratiegui, et al. (2006). "Argonaute slicing is required for heterochromatic silencing and spreading." Science 313(5790): 1134-7.

Jiang, R. H., A. L. Dawe, et al. (2005). "Elicitin genes in Phytophthora infestans are clustered and interspersed with various transposon-like elements." Mol Genet Genomics 273(1): 20-32.

Jones, A. L., C. L. Thomas, et al. (1998). "De novo methylation and co-suppression induced by a cytoplasmically replicating plant RNA virus." EMBO J 17(21): 6385-93.

Jordan, I. K. and J. F. McDonald (1998). "Evidence for the role of recombination in the regulatory evolution of Saccharomyces cerevisiae Ty elements." J Mol Evol 47(1): 14-20.

Jordan, I. K. and J. F. McDonald (1999). "The role of interelement selection in Saccharomyces cerevisiae Ty element evolution." J Mol Evol 49(3): 352-7.

Jordan, I. K. and N. J. Bowen (2004). "Computational analysis of transposable element sequences." Methods Mol Biol 260: 59-71.

Jurka, J. and V. V. Kapitonov (2001). "PIFs meet Tourists and Harbingers: a superfamily reunion." Proc Natl Acad Sci U S A 98(22): 12315-6.

Jurka, J., V. V. Kapitonov, et al. (2005). "Repbase Update, a database of eukaryotic repetitive elements." Cytogenet Genome Res 110(1-4): 462-7.

Kajikawa, M. and N. Okada (2002). "LINEs mobilize SINEs in the eel through a shared 3' sequence." Cell 111(3): 433-44.

Kamau, L., T. Lehmann, et al. (1998). "Microgeographic genetic differentiation of Anopheles gambiae mosquitoes from Asembo Bay, western Kenya: a comparison with Kilifi in coastal Kenya." Am J Trop Med Hyg 58(1): 64-9.

Kaminker, J. S., C. M. Bergman, et al. (2002). "The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective." Genome Biol 3(12): RESEARCH0084.

Kapitonov, V. V. and J. Jurka (2001). "Rolling-circle transposons in eukaryotes." Proc Natl Acad Sci U S A 98(15): 8714-9.

Kapitonov, V. V. and J. Jurka (2003). "Molecular paleontology of transposable elements in the Drosophila melanogaster genome." Proc Natl Acad Sci U S A 100(11): 6569-74.

Kapitonov, V. V. and J. Jurka (2005). "RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons." PLoS Biol 3(6): e181.

Kapitonov, V. V. and J. Jurka (2008). "A universal classification of eukaryotic transposable elements implemented in Repbase." Nat Rev Genet 9(5): 411-2; author reply 414.

Kato, M., K. Takashima, et al. (2004). "Epigenetic control of CACTA transposon mobility in Arabidopsis thaliana." Genetics 168(2): 961-9.

Ke, Z., G. L. Grossman, et al. (1996). "Quetzal: a transposon of the Tc1 family in the mosquito Anopheles albimanus." Genetica 98(2): 141-7.

Kempken, F. and F. Windhofer (2001). "The hAT family: a versatile transposon group common to plants, fungi, animals, and man." Chromosoma 110(1): 1-9.

Kidwell, M. G. and D. Lisch (1997). "Transposable elements as sources of variation in animals and plants." Proc Natl Acad Sci U S A 94(15): 7704-11.

Kidwell, M. G. and D. R. Lisch (2001). "Perspective: transposable elements, parasitic DNA, and genome evolution." Evolution 55(1): 1-24.

Kim, A., C. Terzian, et al. (1994). "Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster." Proc Natl Acad Sci U S A 91(4): 1285-9.

Kim, J. M., S. Vanguri, et al. (1998). "Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence." Genome Res 8(5): 464-78.

Kirkpatrick, M. and N. Barton (2006). "Chromosome inversions, local adaptation and speciation." Genetics 173(1): 419-34.

Klobutcher, L. A. and G. Herrick (1995). "Consensus inverted terminal repeat sequence of Paramecium IESs: resemblance to termini of Tc1-related and Euplotes Tec transposons." Nucleic Acids Res 23(11): 2006-13.

Klobutcher, L. A. and G. Herrick (1997). "Developmental genome reorganization in ciliated protozoa: the transposon link." Prog Nucleic Acid Res Mol Biol 56: 1-62.

Kramerov, D. A. and N. S. Vassetzky (2005). "Short retroposons in eukaryotic genomes." Int Rev Cytol 247: 165-221.

Kristan, M., H. Fleischmann, et al. (2003). "Pyrethroid resistance/susceptibility and differential urban/rural distribution of Anopheles arabiensis and An. gambiae s.s. malaria vectors in Nigeria and Ghana." Med Vet Entomol 17(3): 326-32.

Krzywinski, J., R. C. Wilkerson, et al. (2001). "Toward understanding Anophelinae (Diptera, Culicidae) phylogeny: insights from nuclear single-copy genes and the weight of evidence." Syst Biol 50(4): 540-56.

Krzywinski, J. and N. J. Besansky (2003). "Molecular systematics of Anopheles: from subgenera to subpopulations." Annu Rev Entomol 48: 111-39.

Lai, J., Y. Li, et al. (2005). "Gene movement by Helitron transposons contributes to the haplotype variability of maize." Proc Natl Acad Sci U S A 102(25): 9068-73.

Langley, C. H., E. Montgomery, et al. (1988). "On the role of unequal exchange in the containment of transposable element copy number." Genet Res 52(3): 223-35.

Lehmann, T., W. A. Hawley, et al. (1996). "Genetic differentiation of Anopheles gambiae populations from East and west Africa: comparison of microsatellite and allozyme loci." Heredity 77 ( Pt 2): 192-200.

Lehmann, T., N. J. Besansky, et al. (1997). "Microgeographic structure of Anopheles gambiae in western Kenya based on mtDNA and microsatellite loci." Mol Ecol 6(3): 243-53.

Lehmann, T., W. A. Hawley, et al. (1998). "The effective population size of Anopheles gambiae in Kenya: implications for population structure." Mol Biol Evol 15(3): 264-76.

Lehmann, T., W. A. Hawley, et al. (1999). "The Rift Valley complex as a barrier to gene flow for Anopheles gambiae in Kenya." J Hered 90(6): 613-21.

Lehmann, T., C. R. Blackston, et al. (2000). "The Rift Valley complex as a barrier to gene flow for Anopheles gambiae in Kenya: the mtDNA perspective." J Hered 91(2): 165-8.

Lehmann, T., M. Licht, et al. (2003). "Population Structure of Anopheles gambiae in Africa." J Hered 94(2): 133-47.

Leong, P. T., N. Elissa, et al. (2003). "[Molecular characterization of mosquitoes of the Anopheles gambiae complex from Mayotte and Great Comoro]." Parasite 10(3): 273-6.

Lerat, E., C. Rizzon, et al. (2003). "Sequence divergence within transposable element families in the Drosophila melanogaster genome." Genome Res 13(8): 1889-96.

Levis, R. W., R. Ganesan, et al. (1993). "Transposons in place of telomeric repeats at a Drosophila telomere." Cell 75(6): 1083-93.

Lewis, S. M. and G. E. Wu (1997). "The origins of V(D)J recombination." Cell 88(2): 159-62.

Li, H.-W. (1997). Molecular Evolution, Ed. Sinauer.

Lister, C., D. Jackson, et al. (1993). "Transposon-induced inversion in Antirrhinum modifies nivea gene expression to give a novel flower color pattern under the control of cycloidearadialis." Plant Cell 5(11): 1541-53.

Lorenzi, H. A., G. Robledo, et al. (2006). "The VIPER elements of trypanosomes constitute a novel group of tyrosine recombinase-enconding retrotransposons." Mol Biochem Parasitol 145(2): 184-94.

Lower, R., J. Lower, et al. (1996). "The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences." Proc Natl Acad Sci U S A 93(11): 5177-84.

Luan, D. D., M. H. Korman, et al. (1993). "Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition." Cell 72(4): 595-605.

Lyko, F., B. H. Ramsahoye, et al. (2000). "DNA methylation in Drosophila melanogaster." Nature 408(6812): 538-40.

Mackay, T. F. (1989). "Transposable elements and fitness in Drosophila melanogaster." Genome 31(1): 284-95.

Mackay, T. F., R. F. Lyman, et al. (1992). "Effects of P element insertions on quantitative traits in Drosophila melanogaster." Genetics 130(2): 315-32.

Makalowski, W., G. A. Mitchell, et al. (1994). "Alu sequences in the coding regions of mRNA: a source of protein variability." Trends Genet 10(6): 188-93.

Malik, H. S. and T. H. Eickbush (1999). "Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons." J Virol 73(6): 5186-90.

Manoukis, N. C., J. R. Powell, et al. (2008). "A test of the chromosomal theory of ecotypic speciation in Anopheles gambiae." Proc Natl Acad Sci U S A 105(8): 2940-5.

Marsano, R. M. and R. Caizzi (2005). "A genome-wide screening of BEL-Pao like retrotransposons in Anopheles gambiae by the LTR_STRUC program." Gene 357(2): 115-21.

Martens, J. H., R. J. O'Sullivan, et al. (2005). "The profile of repeat-associated histone lysine methylation states in the mouse epigenome." EMBO J 24(4): 800-12.

Martienssen, R. A., M. Zaratiegui, et al. (2005). "RNA interference and heterochromatin in the fission yeast Schizosaccharomyces pombe." Trends Genet 21(8): 450-6.

Masendu, H. T., R. H. Hunt, et al. (2004). "The sympatric occurrence of two molecular forms of the malaria vector Anopheles gambiae Giles sensu stricto in Kanyemba, in the Zambezi Valley, Zimbabwe." Trans R Soc Trop Med Hyg 98(7): 393-6.

Mathiopoulos, K. D., A. della Torre, et al. (1998). "Cloning of inversion breakpoints in the Anopheles gambiae complex traces a transposable element at the inversion junction." Proc Natl Acad Sci U S A 95(21): 12444-9.

Matzke, M. A., M. F. Mette, et al. (1999). "Host defenses to parasitic sequences and the evolution of epigenetic control mechanisms." Genetica 107(1-3): 271-87.

McClintock, B. (1950). "The origin and behavior of mutable loci in maize." Proc Natl Acad Sci U S A 36(6): 344-55.

McClintock, B. (1956). "Controlling elements and the gene." Cold Spring Harb Symp Quant Biol 21: 197-216.

McClintock, B. (1984). "The significance of responses of the genome to challenge." Science 226(4676): 792-801.

McDonald, J. F. (1993). "Evolution and consequences of transposable elements." Curr Opin Genet Dev 3(6): 855-64.

McLain, D. K., F. H. Collins, et al. (1989). "Microgeographic variation in rDNA intergenic spacers of Anopheles gambiae in western Kenya." Heredity 62 ( Pt 2): 257-64.

McLean, C., A. Bucheton, et al. (1993). "The 5' untranslated region of the I factor, a long interspersed nuclear element-like retrotransposon of Drosophila melanogaster, contains an internal promoter and sequences that regulate expression." Mol Cell Biol 13(2): 1042-50.

McNaughton, J. C., G. Hughes, et al. (1997). "The evolution of an intron: analysis of a long, deletion-prone intron in the human dystrophin gene." Genomics 40(2): 294-304.

Medstrand, P. and D. L. Mager (1998). "Human-specific integrations of the HERV-K endogenous retrovirus family." J Virol 72(12): 9782-7.

Medstrand, P., L. N. van de Lagemaat, et al. (2005). "Impact of transposable elements on the evolution of mammalian gene regulation." Cytogenet Genome Res 110(1-4): 342-52.

Mikkelsen, J. G. and F. S. Pedersen (2000). "Genetic reassortment and patch repair by recombination in retroviruses." J Biomed Sci 7(2): 77-99.

Miller, W. J., A. Nagel, et al. (2000). "Evolutionary dynamics of the SGM transposon family in the Drosophila obscura species group." Mol Biol Evol 17(11): 1597-609.

Montgomery, E., B. Charlesworth, et al. (1987). "A test for the role of natural selection in the stabilization of transposable element copy number in a population of Drosophila melanogaster." Genet Res 49(1): 31-41.

Montgomery, E. A., S. M. Huang, et al. (1991). "Chromosome rearrangement by ectopic recombination in Drosophila melanogaster: genome structure and evolution." Genetics 129(4): 1085-98.

Moran, J. V., R. J. DeBerardinis, et al. (1999). "Exon shuffling by L1 retrotransposition." Science 283(5407): 1530-4.

Mukabayire, O. and N. J. Besansky (1996). "Distribution of T1, Q, Pegasus and mariner transposable elements on the polytene chromosomes of PEST, a standard strain of Anopheles gambiae." Chromosoma 104(8): 585-95.

Mukabayire, O., J. Caridi, et al. (2001). "Patterns of DNA sequence variation in chromosomally recognized taxa of Anopheles gambiae: evidence from rDNA and single-copy loci." Insect Mol Biol 10(1): 33-46.

Muotri, A. R., M. C. Marchetto, et al. (2007). "The necessary junk: new functions for transposable elements." Hum Mol Genet 16 Spec No. 2: R159-67.

Nekrutenko, A. and W. H. Li (2001). "Transposable elements are found in a large number of human protein-coding genes." Trends Genet 17(11): 619-21.

Nene, V., J. R. Wortman, et al. (2007). "Genome sequence of Aedes aegypti, a major arbovirus vector." Science 316(5832): 1718-23.

Neumann, P., D. Pozarkova, et al. (2003). "Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced." Plant Mol Biol 53(3): 399-410.

Noor, M. A., K. L. Grams, et al. (2001). "Chromosomal inversions and the reproductive isolation of species." Proc Natl Acad Sci U S A 98(21): 12084-8.

Omer, S. M. and J. L. Cloudsley-Thompson (1970). "Survival of female Anopheles gambiae Giles through a 9-month dry season in Sudan." Bull World Health Organ 42(2): 319-30.

Onyabe, D. Y. and J. E. Conn (2001). "Genetic differentiation of the malaria vector Anopheles gambiae across Nigeria suggests that selection limits gene flow." Heredity 87(Pt 6): 647-58.

Orgel, L. E. and F. H. Crick (1980). "Selfish DNA: the ultimate parasite." Nature 284(5757): 604-7.

Ostertag, E. M. and H. H. Kazazian (2005). "Genetics: LINEs in mind." Nature 435(7044): 890-1.

Pardue, M. L., O. N. Danilevskaya, et al. (1996). "Drosophila telomeres: new views on chromosome evolution." Trends Genet 12(2): 48-52.

Pearce, S. R., U. Pich, et al. (1996). "The Ty1-copia group retrotransposons of Allium cepa are distributed throughout the chromosomes but are enriched in the terminal heterochromatin." Chromosome Res 4(5): 357-64.

Petrarca, V., J. C. Beier, et al. (1991). "Species composition of the Anopheles gambiae complex (diptera: Culicidae) at two sites in western Kenya." J Med Entomol 28(3): 307-13.

Petrarca, V. and J. C. Beier (1992). "Intraspecific chromosomal polymorphism in the Anopheles gambiae complex as a factor affecting malaria transmission in the Kisumu area of Kenya." Am J Trop Med Hyg 46(2): 229-37.

Petrov, D. A. and D. L. Hartl (1998). "High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups." Mol Biol Evol 15(3): 293-302.

Pimpinelli, S., M. Berloco, et al. (1995). "Transposable elements are stable structural components of Drosophila melanogaster heterochromatin." Proc Natl Acad Sci U S A 92(9): 3804-8.

Popadic, A. and W. W. Anderson (1994). "The history of a genetic system." Proc Natl Acad Sci U S A 91(15): 6819-23.

Powell, J. R., V. Petrarca, et al. (1999). "Population structure, speciation, and introgression in the Anopheles gambiae complex." Parassitologia 41(1-3): 101-13.

Pritham, E. J., C. Feschotte, et al. (2005). "Unexpected diversity and differential success of DNA transposons in four species of entamoeba protozoans." Mol Biol Evol 22(9): 1751-63.

Rizzon, C., G. Marais, et al. (2002). "Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome." Genome Res 12(3): 400-7.

Rohr, C. J., H. Ranson, et al. (2002). "Structure and evolution of mtanga, a retrotransposon actively expressed on the Y chromosome of the African malaria vector Anopheles gambiae." Mol Biol Evol 19(2): 149-62.

Romans, P., R. K. Bhattacharyya, et al. (1998). "Ikirara, a novel transposon family from the malaria vector mosquito, Anopheles gambiae." Insect Mol Biol 7(1): 1-10.

Rozas, J. and M. Aguade (1994). "Gene conversion is involved in the transfer of genetic information between naturally occurring inversions of Drosophila." Proc Natl Acad Sci U S A 91(24): 11517-21.

Rozmahel, R., H. H. Heng, et al. (1997). "Amplification of CFTR exon 9 sequences to multiple locations in the human genome." Genomics 45(3): 554-61.

Rubin, G. M., M. G. Kidwell, et al. (1982). "The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations." Cell 29(3): 987-94.

SanMiguel, P., A. Tikhonov, et al. (1996). "Nested retrotransposons in the intergenic regions of the maize genome." Science 274(5288): 765-8.

SanMiguel, P., B. S. Gaut, et al. (1998). "The paleontology of intergene retrotransposons of maize." Nat Genet 20(1): 43-5.

Sarkar, A., C. Sim, et al. (2003). "Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences." Mol Genet Genomics 270(2): 173-80.

Schiefelbein, J. W., V. Raboy, et al. (1988). "Molecular characterization of suppressor-mutator (Spm)-induced mutations at the bronze-1 locus in maize: the bz-m13 alleles." Basic Life Sci 47: 261-78.

Schlappi, M., R. Raina, et al. (1994). "Epigenetic regulation of the maize Spm transposable element: novel activation of a methylated promoter by TnpA." Cell 77(3): 427-37.

Schneuwly, S., A. Kuroiwa, et al. (1987). "Molecular analysis of the dominant homeotic Antennapedia phenotype." EMBO J 6(1): 201-206.

Seegmiller, A., K. R. Williams, et al. (1996). "Internal eliminated sequences interrupting the Oxytricha 81 locus: allelic divergence, conservation, conversions, and possible transposon origins." Mol Biol Evol 13(10): 1351-62.

Seegmiller, A. and G. Herrick (1998). "A short internal eliminated sequence with central conserved sequences interrupting the LA-MSC gene of the 81 locus in the hypotrichous ciliates Oxytricha fallax and O. trifallax." J Eukaryot Microbiol 45(1): 55-8.

Shao, H. and Z. Tu (2001). "Expanding the diversity of the IS630-Tc1-mariner superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons." Genetics 159(3): 1103-15.

Sharakhov, I. V., A. C. Serazin, et al. (2002). "Inversions and gene order shuffling in Anopheles gambiae and A. funestus." Science 298(5591): 182-5.

Sharakhov, I. V., B. J. White, et al. (2006). "Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in the Anopheles gambiae complex." Proc Natl Acad Sci U S A 103(16): 6258-62.

Sheen, F. M. and R. W. Levis (1994). "Transposition of the LINE-like retrotransposon TART to Drosophila chromosome termini." Proc Natl Acad Sci U S A 91(26): 12510-4.

Sijen, T. and R. H. Plasterk (2003). "Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi." Nature 426(6964): 310-4.

Simard, F., T. Lehmann, et al. (2000). "Persistence of Anopheles arabiensis during the severe dry season conditions in Senegal: an indirect approach using microsatellite loci." Insect Mol Biol 9(5): 467-79.

Singer, M. F., V. Krek, et al. (1993). "LINE-1: a human transposable element." Gene 135(1-2): 183-8.

Slotkin, R. K. and R. Martienssen (2007). "Transposable elements and the epigenetic regulation of the genome." Nat Rev Genet 8(4): 272-85.

Sniegowski, P. D. and B. Charlesworth (1994). "Transposable element numbers in cosmopolitan inversions from a natural population of Drosophila melanogaster." Genetics 137(3): 815-27.

Somboon, P., C. Walton, et al. (2001). "Evidence for a new sibling species of Anopheles minimus from the Ryukyu Archipelago, Japan." J Am Mosq Control Assoc 17(2): 98-113.

Song, S. U., T. Gerasimova, et al. (1994). "An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus." Genes Dev 8(17): 2046-57.

Spanopoulou, E., F. Zaitseva, et al. (1996). "The homeodomain region of Rag-1 reveals the parallel mechanisms of bacterial and V(D)J recombination." Cell 87(2): 263-76.

Stump, A. D., M. Pombi, et al. (2007). "Genetic exchange in 2La inversion heterokaryotypes of Anopheles gambiae." Insect Mol Biol 16(6): 703-9.

Syomin, B. V., L. I. Fedorova, et al. (2001). "The endogenous Drosophila melanogaster retrovirus gypsy can propagate in Drosophila hydei cells." Mol Gen Genet 264(5): 588-94.

Tabara, H., R. J. Hill, et al. (1999). "pos-1 encodes a cytoplasmic zinc-finger protein essential for germline specification in C. elegans." Development 126(1): 1-11.

Takahashi, H., S. Okazaki, et al. (1997). "A new family of site-specific retrotransposons, SART1, is inserted into telomeric repeats of the silkworm, Bombyx mori." Nucleic Acids Res 25(8): 1578-84.

Takahata, N. (1993). "Allelic genealogy and human evolution." Mol Biol Evol 10(1): 2-22.

Taylor, C. E., Y. T. Toure, et al. (1993). "Effective population size and persistence of Anopheles arabiensis during the dry season in west Africa." Med Vet Entomol 7(4): 351-7.

Taylor, C., Y. T. Toure, et al. (2001). "Gene flow among populations of the malaria vector, Anopheles gambiae, in Mali, West Africa." Genetics 157(2): 743-50.

Terzian, C., A. Pelisson, et al. (2001). "Evolution and phylogeny of insect endogenous retroviruses." BMC Evol Biol 1: 3.

Thornburg, B. G., V. Gotea, et al. (2006). "Transposable elements as a significant source of transcription regulating signals." Gene 365: 104-10.

Toure, Y. T., V. Petrarca, et al. (1998). "The distribution and inversion polymorphism of chromosomally recognized taxa of the Anopheles gambiae complex in Mali, West Africa." Parassitologia 40(4): 477-511.

Trelogan, S. A. and S. L. Martin (1995). "Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis." Proc Natl Acad Sci U S A 92(5): 1520-4.

Tripet, F., Y. T. Toure, et al. (2001). "DNA analysis of transferred sperm reveals significant levels of gene flow between molecular forms of Anopheles gambiae." Mol Ecol 10(7): 1725-32.

Tristem, M. (2000). "Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database." J Virol 74(8): 3715-30.

Tubio, J. M., J. C. Costas, et al. (2004). "Evolution of the mdg1 lineage of the Ty3/gypsy group of LTR retrotransposons in Anopheles gambiae." Gene 330: 123-31.

Tubio, J. M., H. Naveira, et al. (2005). "Structural and evolutionary analyses of the Ty3/gypsy group of LTR retrotransposons in the genome of Anopheles gambiae." Mol Biol Evol 22(1): 29-39.

Vanyushin, B. F. (2006). "DNA methylation in plants." Curr Top Microbiol Immunol 301: 67-122.

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science 291(5507): 1304-51.

Vitte, C. and O. Panaud (2003). "Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L." Mol Biol Evol 20(4): 528-40.

Volff, J. N., C. Korting, et al. (2001). "Jule from the fish Xiphophorus is the first complete vertebrate Ty3/Gypsy retrotransposon from the Mag family." Mol Biol Evol 18(2): 101-11.

Volff, J. N. (2006). "Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes." Bioessays 28(9): 913-22.

Voytas, D. F. (1996). "Retroelements in genome organization." Science 274(5288): 737-8.

Walsh, C. P., J. R. Chaillet, et al. (1998). "Transcription of IAP endogenous retroviruses is constrained by cytosine methylation." Nat Genet 20(2): 116-7.

Waterhouse, P. M., M. B. Wang, et al. (2001). "Gene silencing as an adaptive defence against viruses." Nature 411(6839): 834-42.

Waterston, R. H., K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature 420(6915): 520-62.

Weill, M., F. Chandre, et al. (2000). "The kdr mutation occurs in the Mopti form of Anopheles gambiae s.s. through introgression." Insect Mol Biol 9(5): 451-5.

Wessler, S. R. (1988). "Phenotypic diversity mediated by the maize transposable elements Ac and Spm." Science 242(4877): 399-405.

Wicker, T., R. Guyot, et al. (2003). "CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements." Plant Physiol 132(1): 52-63.

Wicker, T., F. Sabot, et al. (2007). "A unified classification system for eukaryotic transposable elements." Nat Rev Genet 8(12): 973-82.

Willis, K. J., L. Gillson, et al. (2004). "Ecology. How 'virgin' is virgin rainforest?" Science 304(5669): 402-3.

Wolffe, A. P. and M. A. Matzke (1999). "Epigenetics: regulation through repression." Science 286(5439): 481-6.

Wondji, C., F. Simard, et al. (2002). "Evidence for genetic differentiation between the molecular forms M and S within the Forest chromosomal form of Anopheles gambiae in an area of sympatry." Insect Mol Biol 11(1): 11-9.

Wondji, C. S., R. H. Hunt, et al. (2005). "An integrated genetic and physical map for the malaria vector Anopheles funestus." Genetics 171(4): 1779-87.

Woodcock, D. M., C. B. Lawler, et al. (1997). "Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon." J Biol Chem 272(12): 7810-6.

Wright, S. (1943). "Isolation by Distance." Genetics 28(2): 114-38.

Wright, S. (1969). The theory of gene frequencies. Evolution and the Genetics of Populations. Chicago, University of Chicago Press. Vol. 2.

Wright, S. (1978). Variability within and among natural populations. Evolution and the Genetics of Populations. Chicago, University of Chicago Press. Vol. 4.

Wright, S. I., N. Agrawal, et al. (2003). "Effects of recombination rate and gene density on transposable element distributions in Arabidopsis thaliana." Genome Res 13(8): 1897-903.

Xiong, Y. and T. H. Eickbush (1990). "Origin and evolution of retroelements based upon their reverse transcriptase sequences." EMBO J 9(10): 3353-62.

Xiong, Y., W. D. Burke, et al. (1993). "Pao, a highly divergent retrotransposable element from Bombyx mori containing long terminal repeats with tandem copies of the putative R region." Nucleic Acids Res 21(9): 2117-23.

Yan, Q., J. Huang, et al. (2003). "Lsh, a modulator of CpG methylation, is crucial for normal histone methylation." EMBO J 22(19): 5154-62.

Yoder, J. A., C. P. Walsh, et al. (1997). "Cytosine methylation and the ecology of intragenomic parasites." Trends Genet 13(8): 335-40.

Zhang, Q., J. Arbuckle, et al. (2000). "Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family Heartbreaker into genic regions of maize." Proc Natl Acad Sci U S A 97(3): 1160-5.

## Agradecimientos

Quiero agradecer a todos aquellos compañeros, amigos y familiares que, de la manera que fuera, han contribuido (a veces sin saberlo) a que pudiera desarrollar mi etapa investigadora, iniciada en el año 2000. En este sentido, quisiera agradecer especialmente la codirección de este trabajo, por su especial preocupación y su contribución muchas veces altruista, al Dr. Emilio Valadé del Río, quien me ha dado la oportunidad de conocer el mundo de la investigación científica y la docencia universitaria. También agradezco a la Dra. Nora J. Besansky y al Dr. Frank F. Collins su invitación para realizar mi estancia de investigación en USA, de la que se han derivado unos exitosos trabajos con los que hasta entonces solo podía soñar pero que, gracias a ellos, son realidad. Gracias al Dr. Horacio Naveira, al Dr. Javier Costas, al Dr. José Luís Bello y al Dr. Antonio Fontdevila, por marcarme muchas veces el camino a seguir y por "adoptarme" fuera de la facultad de Biología de la Universidad de Santiago. Gracias a los compañeros de Hematología por su paciencia y a los compañeros de Medicina molecular, Anatomía patológica, Cardiología y del Departamento de Bioquímica (Grupos del Dr. Jaime Gómez y del Dr. Manuel Rey) por sus conversaciones y el préstamo de sus equipos. Un agradecimiento muy especial a Marta Tojo por su ayuda en el *sprint* final. Gracias al Sr. José Otero por su contribución económica necesaria para realizar el viaje de investigación a USA. Por último, un agradecimiento muy especial a mis padres, en todos los sentidos….

Parte de esta tesis doctoral ha sido financiada con las siguientes ayudas:

Años 2003-2004: beca predoctoral de la Xunta de Galicia y del Fondo Social Europeo (FSE). Duración: 1 año.

Año 2004: beca de estudios en el extranjero de la Xunta de Galicia y el FSE. Duración: 3 meses.

Años 2004-2005: Prórroga beca predoctoral de la Xunta de Galicia y FSE. Duración: 1 año.